


RESEARCH

Open Access



The impact of 24-h take-home exam on language learning and teaching on the China campus of a British university

Xiaomin Ye^{1*} , Yilong Yang¹, Yi Qie¹ and Zengbao Hu²

*Correspondence:
smile.ye@nottingham.edu.cn

¹ Language Centre, University of Nottingham Ningbo China, Ningbo, Zhejiang, China

² School of Economics, University of Nottingham Ningbo China, Ningbo, Zhejiang, China

Abstract

Take-home exam (THE) use has been reported in various disciplines, but research on THE use in language modules in higher education appears to be scarce. The current study employed surveys and interviews to examine how the shift to written THE, in place of the traditional in-class exam (ICE) during the pandemic, impacted language learning and teaching on the China campus of a British university. Additionally, correlation analyses were conducted with ranking data of students from the same cohort under THE and ICE to explore patterns in student performance under these exam conditions. In surveys and interviews, teachers reported that their teaching foci did not change under THE, while many students reported that their learning practices were different under THE and ICE. Students also exhibited a tendency to spend more time practicing skills that they expected to be assessed in the exam. Overall, both teachers and students expressed preference for ICE, with many raising concerns about fairness issues in THE. Furthermore, correlation analyses showed that, overall, for a given group of students, written ICE rankings exhibited strong correlations with each other but written THE rankings did not, suggesting relative instability of THE results. However, when written THE and oral ICE results from the same module are combined, the resultant rankings strongly correlated with pure ICE rankings. This indicates that combining ICE and THE components for assessment could help mitigate some perceived shortcomings, including the instability issue, of THE used alone.

Keywords: Take-home exam, In-class exam, Second language learning and teaching, Exam results, COVID-19

Introduction

Language learning at the university is usually evaluated either through pen and pencil tests, or performance tests (McNamara, 2000). The former is a traditional end-semester final exam, where students are evaluated with “in-class, closed-book, invigilated pen-and-paper exam” (Bengtsson, 2019, p. 1), and is known as in-class exam (ICE). This is a traditional testing method used in many disciplines for a long time with little change (Williams & Wong, 2009). However, the outbreak of COVID-19 has forced educational institutions to change significantly their teaching and assessing methods. Many universities switched from traditional face-to-face teaching to online teaching in a short time and

used new assessment methods. When devising new assessment methods, colleges and universities focused on how to most effectively and authentically assess student learning online (Harrison, 2020). Chan (2022) reviewed the practice of 76 universities and found that the most common approach was to focus on grading, employing for example a “binary grading system” which gives students a “Pass” or “Fail” as a “safe landing” instead of changing the assessment itself, while a few universities, like Princeton University, have replaced all examinations with take-home examinations (p. 8). Take-home exam (THE) is an exam that “the students can do at any location of their choice non-proctored” and whose “time limit is extended to day(s) rather than hours as is the typical time limit for an ICE” (Bengtsson, 2019, p. 2). THE has been used prior to the pandemic as “an assessment method on a regular basis” in universities in Australia, Canada, Finland, and Sweden but was relatively uncommon in UK universities before 2020 (Bone & Maharg, 2019, p. 934), and little investigation has been conducted on its use in the field of second language teaching and learning. Hence, research in this area is necessary.

Context of the current study

The current study was conducted on the Chinese campus of a British university. The campus provides a UK-style education in terms of curriculum, pedagogy, systems, language, and resources (Quality Assurance Agency for Higher Education, 2013), with the same quality assurance standards and regulations as the British main campus. English is the medium of instruction for all subjects except second language modules, where languages are taught in English and the target language. On campus, there are over 8000 students, more than 90% of whom are native Chinese (University of Nottingham Ningbo China [UNNC], n.d.) and share the common language of Mandarin Chinese. They were admitted to the university through the first tier of China’s National College Entrance Examination (Gaokao), before which most of them had not formally studied in a western educational system.

The Chinese students on campus are graduates from domestic high schools nursing an exam-driven learning culture, where “assessment provided motivational forces by offering results indicative of learning progress” for them (Gao, 2006, p. 61). In Chinese high schools, students take English examinations that emphasize the learning of vocabulary and grammar, so students “might develop a belief that learning language is mainly about acquiring knowledge rather than developing communicative skills” (Li & Ruan, 2015, p. 48). After entering our university where the current research was conducted, all the non-English native speaking students take 1 year of English for academic purposes (EAP) courses to support later study. After the preliminary year of English training, students formally start their academic degree study and many of them can choose a second language course in the language center.

The language center (LC) offers French, German, Japanese, Korean, Mandarin, and Spanish courses. In February 2020, when the coronavirus outbreak, the whole university shifted to online teaching. However, in May 2020, the majority of students were able to come back to campus to receive face-to-face teaching while a small number of students unable to come back to campus for various reasons continued learning online in small separate groups. From the beginning of academic year (AY) 2020–2021, because of the uncertainty of COVID, LC announced at the beginning of the autumn semester that the

Table 1 The timeline of THE implementation in LC

	2019–20 spring	2020–21 autumn	2020–21 spring	2021–22 autumn	2021–22 spring
COVID outbreak in China: 28th January 2020. Majority of students came back to campus in May 2020.	Hybrid teaching. Onsite ICE and for those off campus, online-proctored CBE.	Face-to-face teaching. Use of THE was announced at beginning of semester.	Face-to-face teaching. Use of THE was announced at beginning of semester.	Face-to-face teaching. Onsite ICE was reinstated, with an unexecuted contingency plan for THE. Announced at beginning of semester.	Face-to-face teaching. Use of onsite ICE continued, with an unexecuted contingency plan for THE. Announced at beginning of semester.
	Oral exam	Oral exam	Oral exam	Oral exam	Oral exam

Table 2 Structure of assessment at the language center under THE and ICE implementation

Time periods	Component	Weight in final mark	Content and skills	Duration
Academic year 2021–2022	Oral exam	50%	Listening and speaking	20 minutes
	Written exam: in-class exam (ICE)	50%	Reading (40%); Use of language (grammar) (10%); Writing (50%) – one piece of writing based on reading stimulus;	1.5 to 2 hours depending on level
Academic year 2020–2021	Oral exam	50%	Listening and speaking	20 minutes
	Written exam: 24-hour take-home exam (THE)	50%	Writing (100%) – two pieces of writing based on reading stimulus;	24 hours

In academic year 2020–2021, the usual 1.5- to 2-h in-class written exam was replaced by 24-h THE and the exam tasks were changed to two pieces of writing based on two reading stimuli. Regardless of format, the written exam weighed 50% in the final mark of a module. The oral exam (always an ICE) constituted the other 50%

end-semester written exam would change to online entirely with a 24-h THE and the use of THE format lasted for 2 semesters until it changed back to ICE in autumn 2021 (see Table 1).

The implementation of THE

When transferring to online format, in autumn 2020, the traditional 1.5- to 2-h in-class written exam was replaced by a 24-h THE and the exam tasks were changed to two pieces of writing based on two reading stimuli (see Table 2). In the 24-h THE, students would access Moodle, where the module convener had set up the THE exam assignment before the exam date. Students would download the exam paper and were expected to upload their answers onto Moodle within 24 h. We chose THE as the end-semester assessment for several reasons. Firstly, not all students had been able to come back to campus, so it was not possible to resume the traditional ICE. Secondly, THE was one of the alternative methods employed by institutions in other parts of the world where teaching was still totally online (Gamage et al., 2020), including the LC on our home campus in the UK. Thirdly, for many researchers, THE was “a promising move to assessment for learning during the time of Covid-19” (Tam, 2022, p. 488) and could even be “far more valuable than being an emergency alternative to in-class exam” (Braselmann

et al., 2022, p. 99). Finally, it should be noted that we gave students 24 h to finish their exam considering that some students were outside of China and had a time difference, and that technical issues could arise and take time to resolve. The LC considered also the potential problems related to THE, namely that students could check books, internet, and so on for answers. To cope with these problems, we changed the exam design. The regular in-class written exam consisted of three parts: reading comprehension, use of language (grammar), and writing, whereas in the 24-h THE, there were two writing tasks based on two reading stimuli, where students were required to read the stimulus text and write personal responses. The reading stimuli were provided in picture format so students could not directly copy and paste any text, in the hope that this could reduce the potential use of tools like machine translation.

Students were asked to hand-write their answers either on writing sheets centrally designed by LC or on blank white sheets of paper. They would then scan or take pictures of their work and upload the files on Moodle. After the end of the 24-h period, module conveners downloaded all students' scripts and distributed them within their respective language team for marking. The marking criteria were centrally designed by the home university, but there was no time to discuss them and standardize their use across language teams on our campus, so standardization was conducted only within each language team.

At our university, language modules also include an oral exam, whose format did not change during the pandemic. The oral exam is normally an in-class closed-book exam, where groups of 2 or 3 students engage in a conversation prompted by a randomly drawn card, after which each student answers some questions asked by examiners on the spot. For the oral exam, the only change caused by the pandemic was that a very small number of students had to be examined online with the same invigilation and procedural standards. For the current study, therefore, the oral exam is conceptualized as always being an ICE.

Literature review

In the literature, there are two other concepts related to THE and ICE and appear frequently in research studies: open-book exams (OBEs) and closed-book exams (CBEs). It is important to first make clear the usage of these terms for the purposes of the current study. First, in OBEs, students can access class notes and teaching materials during the exam (e.g., Tao & Li, 2012). OBEs could be in-class invigilated open-book exams or take-home open-book exam. The latter happens in the same condition as THE. In this paper, therefore, the term THE is used for all the exams that students take at home or other places without invigilators during a given period of time, from a few hours to a few days; these include the take-home OBEs. Second, closed-book exams (CBEs) usually are in-class exams (ICEs), so in this paper, ICEs and CBEs are considered interchangeable and are used to refer to exams that happen in class with invigilators where students are not allowed to access learning materials. Closed-book take-home exams (e.g., Fernald & Webster, 1991) are not considered for the current study.

THE have been used and studied in various disciplines, such as psychology (Rich, 2011), social sciences (Spiegel & Nivette, 2021), education (Braselmann et al., 2022; Şenel & Şenel, 2021), computer sciences (López et al., 2011), chemistry (Clark et al.,

2020; Jacobs, 2021; Raje & Stitzel, 2020), nursing (Tao & Li, 2012), medical science (Ng, 2020), and law (Bone & Maharg, 2019), and some were conducted with participants from mixed disciplines (Karagiannopoulou & Milienos, 2013; Marsh, 1984; Tam, 2022; Williams & Wong, 2009).

There are also review reports. Bengtsson (2019) reviewed 35 articles about THE in higher education and concluded that THE may be the preferred choice of assessment because they promote higher-order thinking skills and allow time for reflection. Durning et al. (2016) and Johanns et al. (2017) have done similar work.

However, there seems to be a dearth of research on the use of THE in second language education.

Impact of THE on learning and teaching

Researchers reported inconsistent results about the impact of THE (including take-home OBEs) on students' learning.

Some researchers found that THE had a positive impact on students' learning. For example, López et al. (2011) concluded that THE is "a powerful tool" for assessing all types of skills (p. 6) and was greatly appreciated by students as it improved their learning process. Later studies also echoed these statements, reporting that THE implementation deepened understanding (Karagiannopoulou & Milienos, 2013; Jacobs, 2021), and increased student motivation and engagement (Myry & Joutsenvirta, 2015).

These results, however, must be viewed with care, as these benefits might be associated with changes in other factors brought about by THE implementation, such as changes in question types, rather than the shift to THE per se. For instance, in López et al. (2011), more than half of the student participants found their THE in a computer science course "very demanding" and agreed that significant learning took place during the exam (pp. 5–6). This could be due to the fact that López and colleagues designed a THE with 16 open-ended questions that required students to collect and synthesize information from the course and on the internet (p. 5). Similarly, when shifting to THE, Jacobs (2021) made more extensive use of open-ended questions in chemistry examinations. Another example is Braselmann and colleagues' (2022) study: the researchers did not only create a complex THE design with a mixture of closed, semi-open, and open-ended components, but they redesigned the whole course around the THE requirements. When reviewing the reported benefits of THE, it is therefore essential to identify and consider the factors affected by THE use in individual studies.

While the abovementioned THE studies involved certain degrees of redesign of exam tasks, Marsh (1984) gave two groups of students the same set of multiple-choice questions in THE and ICE formats, respectively. In an unexpected delayed test 1 week later, the students in the ICE group outperformed those in the THE group, indicating that ICE may be associated with better retention compared with THE. As Marsh explained, students' expectation that they could rely on study materials in a THE could hinder learning (p. 112).

None of the studies discussed so far was from the field of language education, where exams are designed to assess language knowledge and skills and can look very different

from exams in other disciplines. Thus, the current study contributes to the discussion on THE's impact on learning and teaching by examining its use in language education.

Exam results

Extant studies that compared student results in THE and ICE mainly used average scores or grades in the same course. For example, Braselmann and colleagues (2022) compared average grades under THE with previous ICE grades in the same course and found “no significant difference” (p. 97). Spiegel and Nivette (2021) also reported comparable results between THE and ICE. In Jacobs (2021), average scores were overall higher for THE than ICE, but not by much.

These comparisons have some limitations. First, given the relatively small amount of data, no pattern can be reliably observed yet and the comparisons do not reveal much about the exam formats. Second, some of these studies (e.g., Braselmann et al., 2022) compared results from different cohorts. This may make results harder to interpret, as different cohorts may possess varying characteristics that affect exam results, and therefore may also limit the insight that can be gained through comparing THE and ICE results.

In response to these limitations, the current study utilized marks from the same cohort across four consecutive semesters, as well as the corresponding ranking information, to explore patterns in performance of the same students in THE and ICE.

Exam design and fairness

The majority of research produced about THE and ICE has demonstrated that, either because of its format (open-ended and essay-type questions) or because of the longer time duration, THE promotes high-order cognitive activities (Bengtsson, 2019; Tam, 2022), evaluation and creation of knowledge (Khan, 2022), and reflection on personal experience (Ng, 2020).

However, a basic principle in assessment design is to ensure that assessments enable students to demonstrate their learning (Quality Assurance Agency for Higher Education, 2018), and “reflect students' real competence” (Şenel & Şenel, 2021, p. 246). The accuracy of the assessment in representing students' real competence not only depends on effective design of exam tasks but is also related to students' conduct during exams; students might misbehave in exams and violate academic integrity, which can impact exam fairness. Because of the open-book nature of THE, dishonesty is a common concern (Cleophas et al., 2021; Ng, 2020). Studies have revealed that there is a higher probability for online students to cheat in assessments in comparison with campus-based students (e.g., Gamage et al., 2020).

Another consideration is, when adopting THE, it is also necessary to consider the overall assessment design of the module. Durning et al. (2016) and Johanns et al. (2017) concluded that a combined approach (of OBEs/THE or CBEs/ICEs) could be more effective in assessing different competencies. The current study will examine this point in its data analysis.

All the research discussed so far focused on content modules in non-language subjects, and no research was found related to language examinations, which often aim at

enabling students to demonstrate the mastery of multidimensional skills in a language rather than knowledge in a subject. This study also contributes to filling this gap.

Research questions

The research discussed so far focused on THE as an assessment tool in content modules aiming at testing students' knowledge of concepts and theories and their ability to apply them in disciplinary contexts. Most studies that analyzed exam results did so only in terms of average marks or grades from different cohorts. Furthermore, few studies surveyed both students' and teachers' perspectives on THE and ICE, as well as their learning and teaching strategies under these exam formats.

Given these gaps, the current study aims at contributing to the ongoing discussion about THE by exploring its use in the field of language education, and more specifically by answering the following research questions:

1. To what extent does the awareness of take-home exam (THE) implementation affect students' learning strategies and teachers' teaching strategies?
2. How well correlated are exam results of a given cohort under THE and ICE as measured by rankings?
3. How do students and teachers believe THE should be designed to assess language skills accurately?

Material and methods

Methods and samples

A combination of quantitative and qualitative data collection, including surveys and semi-structured interviews, was adopted in this study. The findings were triangulated with the analysis of exam results.

We conducted two questionnaire surveys to gather information from students and teachers respectively about their behaviors, opinions, and experiences in relation to THE, and explored these data for possible similarities and differences (Neuman, 2014). As Wellington (2015) posed it, a survey is a "fact-finding mission" (p. 191); this reflects the purpose of our two surveys. In addition, to gain a deeper, more nuanced understanding of the participants' survey answers and their subjective experience with THE (Kvale, 2008), respondents were also invited to an interview.

After obtaining ethical approval from the university, in May 2022, two anonymous surveys were launched to both language center students and teachers respectively. The surveys were sent to all 23 language center teachers and all 424 students who took A2-B1 level language modules in 2021–2022. We targeted this cohort of students because they experienced both the 24-h THE in 2020–2021 and the ICEs in 2021–2022 and therefore could compare their THE and ICE experience.

Each survey consisted of three main parts. For students, the first part included questions on basic background information, including the language they were learning and the level of their language module at LC. The second part focused on their learning strategies and practices under THE and ICE, and their attitudes toward the two assessment formats. In the third part, students were invited to present their opinions

on preferred exam format and question types. The last two parts included both Likert scale and open-ended questions to enable respondents to elaborate on their answers. For language teachers, the survey questions covered their teaching strategies under THE and ICE, their attitudes toward the two exam formats, and what they thought should be taken into consideration when designing THE. This survey also contained Likert scale and open-ended questions.

Before officially launching the survey, a pilot survey was conducted with five students and two language teachers. The surveys were then refined following their feedback for higher effectiveness and reliability. Changes were made (1) to reorder some questions; (2) in phrasing to improve readability; (3) in the number of options for closed questions; and (4) to correct typos.

Out of the 424 eligible students, 135 completed the survey. Table 3 displays their demographic information. French (32.6%), Japanese (30.4%), and Spanish (25.2%) were the three languages with the most respondents. Out of the 23 teachers, 11 completed the survey.

At the end of the survey, nine students and five teachers indicated interest in being interviewed. Interviews lasted about 20 min each for students and 30–45 min each for teachers. Interview recordings were transcribed after each interview.

In addition, to triangulate our findings, quantitative analysis was conducted to compare the THE results in 2020–2021 and ICE results in 2021–2022. We collected the exam results of 206 students, of whom 28.3% were studying French, 25.9% Japanese, and 45.9% Spanish.

Data analysis

We adopted four empirical methods for quantitative analysis: descriptive analysis, scale analysis, non-parametric analysis, and correlation analysis, as illustrated in Fig. 1. Descriptive analysis revealed important facts and patterns of respondents' language learning and teaching practices under different assessment methods. Scale analysis enabled us to measure respondent attitudes toward different assessment methods, and included both validity (Kaiser-Meyer-Olkin [KMO] test and Bartlett's test; Bartlett, 1954; Dabestani et al., 2014; Gunawan et al., 2022; Kaiser, 1974) and reliability tests (Cronbach's alpha coefficient; Croasmun & Ostrom, 2011; Gliem & Gliem, 2003; Vaske et al., 2017), both of which are commonly used methods to test the reliability and validity of Likert-scale questionnaires. Our survey included rank-order questions where

Table 3 Demographic characteristics of students

Variable	Option	Obs	Percent
Language	French	44	32.6%
	German	7	5.2%
	Japanese	41	30.4%
	Korean	2	1.5%
	Mandarin	7	5.2%
	Spanish	34	25.2%
Total		135	

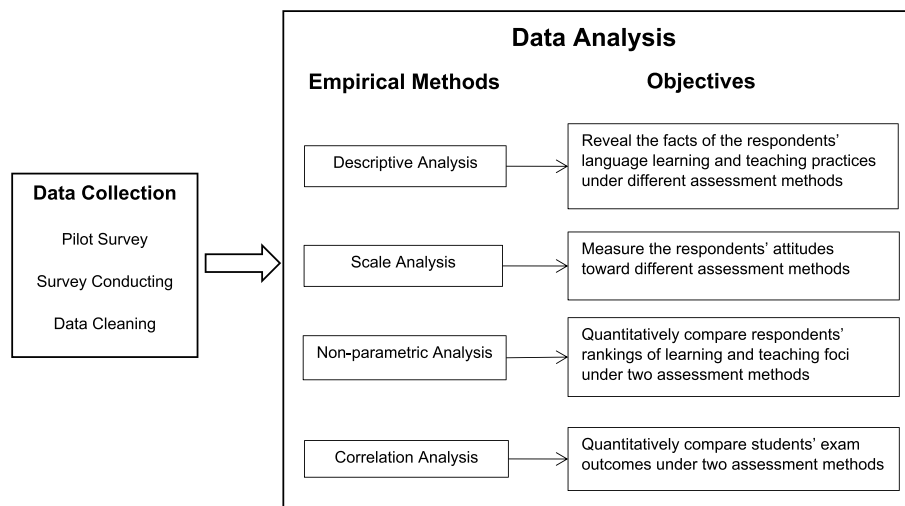


Fig. 1 Approaches and methods of quantitative analysis

respondents were asked to rank their learning or teaching foci under the two assessment methods, and non-parametric tests can compare these ranked data more robustly than parametric tests (Krzywinski & Altman, 2014). Hence, we followed Shin and Park's (2009) model and used the non-parametric Wilcoxon signed rank test to compare learning and teaching foci under the two assessment formats. As for analyses of student exam results, we conducted paired *t*-tests and the Spearman correlation test to compare results under THE and ICE. For comparison of cohort average marks under THE and ICE, we employed the paired *t*-test because several previous simulations have found parametric tests to be more robust in analyzing both normally and non-normally distributed continuous data in most situations when the sample size is not very small (Skovlund & Fenstad, 2001; Wadgave & Kahairnar, 2019). Moreover, to gain further insight into how students performed under different exam settings, it is vital to also analyze changes in students' relative positions within their cohort, as "rank eliminates any disparity between the two characteristics compared" (Spearman, 2010, p. 1141). Therefore, Spearman's rank correlation coefficient analysis, one of the most known tests for comparing rankings (Csató, 2013), was conducted to compare students' exam rankings.

Regarding qualitative analysis, to ensure interrater reliability, we followed the following steps. Two researchers of the team coded two students' and one teacher's interview transcripts separately using *descriptive codes* (Saldaña, 2013). They then met, and selected and retained the codes that they had both identified or deemed relevant to the study, based on which they built the codebook (see Table 4). Afterwards, one of the researchers coded all the students' and teachers' interviews including those already coded (in total 11 students and 5 teachers). NVivo 12 was used to organize and manage the codes.

Table 4 Codebook

Codes	Sub-codes
THE impact on language learning/teaching	<ul style="list-style-type: none"> • Students’ learning practices during the semester/when preparing for the exam/during the exam period under different exam conditions (ICE/THE) • Teachers’ teaching methods during the semester/when preparing students for the exam/during the exam period under different exam conditions (ICE/THE) • Students’ perceived level of psychological wellbeing under different exam conditions (ICE/THE)
Difference between students’ performance in THE and ICE as measured by exam results	<ul style="list-style-type: none"> • Students’/teachers’ understanding and interpretation of exam results • Students’/teachers’ explanation of differences in exam results • Perceived factors influencing exam results
Considerations about exam design	<ul style="list-style-type: none"> • Students/teachers’ opinions about THE design • The ideal written exam according to students’/teachers’ belief • The ideal assessment format to optimize written exam effectiveness • What type of curriculum design ensures assessment effectiveness • Exam paper design and academic misconduct

Table 5 Impact of THE on students’ out-of-class learning

Impact of THE	Statement	Obs	Percent	Mean	Std. Dev
Change in time of self-study each week ^a	I spent fewer hours	35	26.7%	1.82	0.57
	I spent more hours	12	9.2%		
	No Impact	84	64.1%		
	Total	131			
Change in learning focus on language skills ^b	No	75	55.6%	0.44	0.50
	Yes	60	44.4%		
	Total	135			
Change in extracurricular practices ^c	No	37	28%	0.72	0.45
	Yes	95	72%		
	Total	132			

^a 1=fewer hours, 2=no impact, 3=more hours (self-study includes any study activity outside of class time.)

^b 0=no, 1=yes (learning foci include vocabulary, grammar, reading, and writing.)

^c 0=no, 1=yes (extracurricular practices include such activities as consuming recreational contents in the target language, participating in events organized by language teachers, following off-campus language classes, and using language apps.)

Results

In this part, we first present the quantitative results based on the surveys and the exam results and then present the qualitative results from the interviews and the open-ended questions in the surveys. In both parts, we present results in the following topical order. *Impact of THE on Learning and Teaching, Exam Results, and Exam Design and Fairness.*

Quantitative analysis results: surveys and THE-ICE result comparison

Impact of THE on learning and teaching

The student survey shows that most students believe that THE did not influence their class attendance (83.05%), class participation (82.76%), and homework completion (76.72%).

Table 6 Comparisons of students' self-study time and focus for THE and ICE

Pair	Assessment method	Skill	Obs	Mean ^a	Std. Dev	Paired differences in mean	t-statistic	p-value
Pair 1	THE	Vocabulary	98	3.33	1.50	-0.10	-0.544	0.558
	ICE			3.43	1.36			
Pair 2	THE	Grammar	99	3.65	1.58	-0.13	-0.713	0.478
	ICE			3.78	1.37			
Pair 3	THE	Reading	97	2.59	1.46	-0.43	-2.359	0.020
	ICE			3.02	1.43			
Pair 4	THE	Writing	98	3.24	1.58	0.14	0.667	0.506
	ICE			3.10	1.53			

^a Estimated number of hours of study per week

Table 7 Teachers' teaching focus for ICE and THE

Descriptive statistics			
Assessment method	Teaching focus	Mean rank ^a	Obs
ICE	Vocabulary	4.38	8
	Grammar	2.75	8
	Phonetics	6.38	8
	Listening	4.25	8
	Speaking	2.88	8
	Reading	3.88	8
	Writing	3.50	8
THE	Vocabulary	4.00	8
	Grammar	2.75	8
	Phonetics	6.13	8
	Listening	4.63	8
	Speaking	3.00	8
	Reading	4.25	8
	Writing	3.25	8

^a Respondents were required to rank the options from the most important (1) to the least (7)

Table 5 shows that the majority of students reported that they did not change their weekly self-study time (64.1%). Compared with ICE, 26.7% reported spending less time on self-study for THE, while 9.2% spent more time for THE. Most students did not change their self-study focus (55.6%). According to Table 6, grammar (mean=3.65, SD=1.58 for THE; mean=3.78, SD=1.37 for ICE) and vocabulary (mean=3.33, SD=1.50 for THE; mean=3.43, SD=1.36 for ICE) were the two aspects student spent the most time on, regardless of assessment type. In contrast, reading was the skill they spent the least time on for both types of assessment (mean=2.59, SD=1.46 for THE, and mean=3.02, SD=1.43 for ICE). Students spent significantly more time on reading-focused self-study under ICE implementation compared with THE (paired differences in mean=-0.43, $t=-2.359$, $p=0.02$).

Table 7 shows that teachers (Obs=8) focused on grammar (mean rank=2.75 for THE and 2.88 for ICE) and speaking (mean rank=3.00 for THE and 2.88 for ICE) skills the most for both THE and ICE. According to the Wilcoxon signed rank test shown

Table 8 Comparison of teaching foci under THE and ICE

THE - ICE		Obs	Mean rank	Sum of ranks	Wilcoxon signed ranks test	
					Z	Asymp. Sig
Vocabulary	Negative ranks	2	2.25	4.50	−0.816 ^a	0.414
	Positive ranks	1	1.50	1.50		
	Ties	5				
Grammar	Negative ranks	1	3.00	3.00	0.000 ^b	1.000
	Positive ranks	2	1.50	3.00		
	Ties	5				
Phonetics	Negative ranks	2	1.50	3.00	−1.414 ^a	0.157
	Positive ranks	0	0.00	0.00		
	Ties	6				
Listening	Negative ranks	1	1.50	1.50	−0.816 ^c	0.414
	Positive ranks	2	2.25	4.50		
	Ties	5				
Speaking	Negative ranks	1	2.00	2.00	−0.577 ^c	0.564
	Positive ranks	2	2.00	4.00		
	Ties	5				
Reading	Negative ranks	1	2.00	2.00	−1.134 ^c	0.257
	Positive ranks	3	2.67	8.00		
	Ties	4				
Writing	Negative ranks	2	4.25	8.50	−0.272 ^a	0.785
	Positive ranks	3	2.17	6.50		
	Ties	3				

^a Based on positive ranks

^b The sum of negative ranks equals the sum of positive ranks

^c Based on negative ranks

Table 9 Impact of THE on language teachers' teaching activities

Activity	Obs	Mean ^a	Std. Dev
Giving more writing tasks for homework	9	3.22	1.39
Giving extra training on essay writing	9	3.22	1.39
Introducing online resources on writing skills	9	3.11	1.54
Reducing homework or online activities related to reading	9	2.25	1.58

^a 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree

in Table 8, there was no statistically significant change in teachers' teaching focus on vocabulary, grammar, phonetics, listening, reading, and writing when THE was implemented. Teachers' teaching activities did not change for THE either, as shown by their overall neutral or "disagreeing" responses to questions about the impact of THE on their teaching practices (Table 9).

Students' attitudes toward 24-h take-home exams were measured in terms of perceived benefits, issues, and level of stress, while their attitudes toward traditional exams were measured in terms of perceived benefits and level of stress, all on a 5-point Likert scale. Table 10 and Table 11 provide results of the reliability and validity check. As shown in Table 10, the Cronbach's alpha of all items are greater than 0.5, indicating good

Table 10 Reliability check of students' attitudes toward different assessment methods

Item		Mean	Obs ^a	Corrected item-total correlation	Squared multiple correlation	Cronbach's alpha if item deleted	Cronbach's alpha based on standardized items
Benefits of THE	More time to write and check	4.63	118	0.318	0.153	0.461	0.562
	Less pressure because there is more time	4.44	118	0.381	0.365	0.385	
	Less pressure because there is no invigilator	3.82	118	0.486	0.33	0.244	
	Chance to seek help	3.38	118	0.149	0.083	0.636	
Issues of THE	Spending less time on learning	2.77	82	0.626	0.45	0.75	0.800
	Not doing homework unless related to exam	2.15	82	0.56	0.374	0.77	
	Reading and grammar skills are less developed	3.01	82	0.676	0.471	0.731	
	24hrs is too long	2.89	82	0.598	0.427	0.758	
	Technical issues	2.30	82	0.465	0.262	0.796	
Stress of THE	Time constraint	1.96	77	0.447	0.247	0.655	0.698
	Not good at writing	2.82	77	0.386	0.252	0.694	
	Fear of technical issues	2.05	77	0.516	0.475	0.607	
	Fear of exam conditions	2.35	77	0.595	0.502	0.548	
Benefits of ICE	Fair for everyone	3.95	120	0.753	0.594	0.848	0.883
	Demonstrating my skills	3.50	120	0.825	0.682	0.785	
	Help me to obtain higher marks	3.28	120	0.741	0.565	0.863	
Stress of ICE	Time constraint	4.27	130	0.634	0.402	-	0.776
	Exam environmental pressure	3.92	130	0.634	0.402	-	

^a Listwise deletion based on all variables in the procedure

internal consistency; hence, the questionnaire has good reliability in measuring respondent attitudes (Vaske et al., 2017). In the reliability test, an observation should be dropped if it has a missing value in at least one of the specified variables, so the numbers of observations in Table 10 might be different from those in Table 12.

Table 11 Validity check of students' attitudes toward different assessment methods

Item	Kaiser-Meyer-Olkin measure of sampling adequacy	Bartlett's test of sphericity		
		Approx. chi-square	df	sig
24hr take-home exam (THE)	0.721	278.739	78	<0.001
Traditional exam (ICE)	0.723	271.008	10	<0.001

Table 12 Students' opinions about THE and ICE

Item		Obs	Mean ^a	Std. Dev
Benefits of THE	More time to write and check	131	4.58	0.77
	Less pressure because there is more time	131	4.40	1.02
	Less pressure because there is no invigilator	132	3.80	1.23
	Chance to seek help	121	3.38	1.45
Issues of THE	Spending less time on learning	109	2.58	1.34
	Not doing homework unless related to exam	101	1.99	1.22
	Reading and grammar skills are less developed	108	2.73	1.50
	24hrs is too long	102	2.91	1.41
	Technical issues	99	2.34	1.29
Stress of THE	Time constraint	92	1.95	0.92
	Not good at writing	113	2.83	1.19
	Fear of technical issues	100	2.14	1.11
	Fear of exam conditions	100	2.47	1.25
Benefits of ICE	Fair for everyone	130	3.87	1.31
	Demonstrating my skills	128	3.43	1.30
	Help me to obtain higher marks	123	3.24	1.41
Stress of ICE	Time constraint	132	4.27	0.90
	Exam environmental pressure	130	3.92	1.12

^a 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree

According to Table 11, the result of the KMO test for 24-h take-home exams is 0.721 ($p < 0.001$), and that for traditional exams is 0.723 ($p < 0.001$), meaning that the questionnaire has good validity in measuring respondent attitudes (Kaiser, 1974).

As shown in Table 12, students thought that the exam duration of THE was too long, that they developed less in reading and grammar skills, and that they spent less time on learning overall. Similarly, teachers thought that, under THE implementation, students' reading and grammar skills were less developed, and that students spent less time on language learning and did not do their homework (Table 13).

Exam results

The comparison of exam results of ICEs and THE (Table 14) shows that during the pandemic, in 2021–2022, there was a statically significant drop in the average mark for French (from 64.4 to 54.4 out of 100 marks) and Spanish (from 61.6 to 58.6), while the Japanese average mark increased (from 59.6 to 61.2).

Spearman's rank correlation coefficient analysis was conducted to compare students' rankings within their cohort under different exam conditions. First, we compared the ranking changes of written exam and oral exam in the two semesters within 2020–2021

Table 13 Teachers’ opinions about THE and ICE

Item		Obs	Mean ^a	Std. Dev
Benefits of THE	It evaluates student’s deep-thinking skills	4	2.75	0.96
	Students have less pressure	4	4.75	0.50
	Students have more time to write and check answers	4	4.50	0.58
	Students have the chance to seek help	4	4.50	1.00
Issues of THE	Suspected plagiarism	11	4.45	0.82
	Students spend less time on language learning	11	3.27	1.01
	Students don’t do other homework	11	2.91	1.04
	Students’ reading and grammar skills are less developed	11	3.18	1.08
	24hrs is too long	11	3.60	0.97
	It created more workload for the convenor	11	4.00	0.89
Benefits of ICE	Fair for everyone	11	4.73	0.47
	Demonstrating students’ skills	11	3.55	1.29
	Help me to obtain higher marks	11	3.70	0.95

^a 1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree

Table 14 Comparisons of average marks in written THE and ICEs

Language	Academic year	Assessment method	Average mark	Paired differences in mean	p-value
French	20–21 academic year	THEs	64.39	9.97	<0.001
	21–22 academic year	ICEs	54.42		
Japanese	20–21 academic year	THEs	59.62	−1.58	0.078
	21–22 academic year	ICEs	61.20		
Spanish	20–21 academic year	THEs	61.63	3.05	0.016
	21–22 academic year	ICEs	58.58		

Table 15 Comparisons of students’ rankings in different semesters of the same academic year

Language	Comparing factors	Written exam ranking		Oral exam ranking		n
		Spearman correlation coefficient	Sig. (2-tailed)	Spearman correlation coefficient	Sig. (2-tailed)	
French	AY2020–2021 Sem A	0.541** (THE)	<0.001	0.697**	<0.001	55
	AY2020–2021 Sem B					
	AY2021–2022 Sem A	0.712** (ICE)	<0.001	0.748**	<0.001	
	AY2021–2022 Sem B					
Japanese	AY2020–2021 Sem A	0.399** (THE)	0.003	0.688**	<0.001	53
	AY2020–2021 Sem B					
	AY2021–2022 Sem A	0.636** (ICE)	<0.001	0.802**	<0.001	
	AY2021–2022 Sem B					
Spanish	AY2020–2021 Sem A	0.422** (THE)	<0.001	0.625**	<0.001	91
	AY2020–2021 Sem B					
	AY2021–2022 Sem A	0.773** (ICE)	<0.001	0.689**	<0.001	
	AY2021–2022 Sem B					

and 2021–2022 (Table 15). The comparison shows that, for each of the three languages, the oral exam rankings were strongly correlated with each other, with ρ values between 0.625 and 0.802 ($p < 0.001$). On the other hand, the written exam rankings were not strongly correlated in 2020–2021 when THE was implemented (ρ between 0.399 and 0.541, $p \leq 0.003$), while they exhibited much stronger correlations in 2021–2022 under ICE (ρ between 0.636 and 0.773, $p < 0.001$).

The written mark, regardless of THE or ICE, only consists of 50% of the final mark of the language modules, so we then included the oral exam (always ICE), the other 50%, into our analysis.

We used rankings obtained from average marks between semesters within the same academic year to indicate overall results in that academic year. We compared students' rankings for both written and oral components in 2020–2021 (with a written THE) and 2021–2022 (with a written ICE, Table 16). The results show that the correlations for the written exams were weak for French ($\rho = 0.331$, $p = 0.014$) and Spanish ($\rho = 0.346$, $p < 0.001$) but strong for Japanese ($\rho = 0.699$, $p < 0.001$); the correlations for the oral component for all languages were strong, between 0.695 and 0.770 ($p < 0.001$).

When we used the rankings obtained by combining the oral and written components (i.e., based on the final overall module results; written 50%, oral 50%), correlation analyses showed strong correlations between 2020–2021 and 2021–2022 rankings for each of the three languages (ρ between 0.682 and 0.787, $p < 0.001$), despite the inconsistency in results observed when the written component alone was used for analysis (Table 16).

The surveys also yielded relevant information in understanding student results under THE and ICE. For THE, students reported (Table 12) that they had more time to write and check their answers (mean = 4.58, SD = 0.77) and that they felt less pressure because there was more time and no invigilator and they had the chance to seek help. However, some thought that 24 h was too long for THE (mean = 2.91, SD = 1.41). On the

Table 16 Comparisons of students' rankings from written, oral, and overall (written and oral 50% each) marks

Comparing factors	French			Japanese			Spanish		
	Spearman correlation coefficient	Sig. (2-tailed)	<i>n</i>	Spearman correlation coefficient	Sig. (2-tailed)	<i>n</i>	Spearman correlation coefficient	Sig. (2-tailed)	<i>n</i>
AY2020–2021 written (THE) AY2021–2022 written (ICE)	0.331*	0.014	55	0.699**	<0.001	53	0.346*	<0.001	91
AY2020–2021 oral AY2021–2022 oral	0.695**	<0.001		0.759**	<0.001		0.770**	<0.001	
AY2020–2021 overall AY2021–2022 overall	0.685**	<0.001		0.787**	<0.001		0.682**	<0.001	

other hand, students believed that ICEs were fair for everyone (mean=3.87, SD=1.31) and allowed them to demonstrate their skills (mean=3.43, SD=1.30) and obtain higher marks (mean=3.24, SD=1.41). According to the teacher respondents (Table 13), the major issue of THE was suspected plagiarism (mean=4.45, SD=0.82), followed by the possibility of technical issues (mean=3.64, SD=1.12). Some also found the exam duration too long (mean=3.60, SD=0.97). Like students, teachers believed that ICEs were fair for everyone (mean=4.73, SD=0.47) and enabled students to obtain higher marks (mean=3.70, SD=0.95) and to demonstrate their skills (mean=3.55, SD=1.29).

In the interviews, students and teachers also gave comments that could help interpret the result differences under the two exam formats. These will be presented in “[Exam results](#)”.

Exam design and fairness

Based on their experience with THE, survey respondents were asked to express their preference for different options for THE exam design in relation to such aspects as exam content, exam duration, learning skills, and learning focus required when preparing for THE.

For THE content, about half of the student respondents (46.5%) preferred a comprehensive exam consisting of reading, use of language, and one writing task; 27.9% preferred two writing tasks, while 25.6% preferred one reading task and one writing task. For the exam duration, 44.4% of the student respondents preferred a take-home exam with a 4-h time limit, while 31.6% preferred a 24-h take-home exam. Most students believed that a good THE design should require them to use critical thinking and resource finding abilities (60.15%), time management skill (52.63%), and high-order thinking (34.59%), and require their learning foci to be writing (74.44%) and grammar (57.89%).

On the other hand, in the eyes of the teacher respondents, a good THE should be only focusing on writing tasks (70%) and should have a time limit of 2 h (70%). A good design should require students to utilize time management skills (mean=3.91, SD=1.14), critical thinking (mean=3.82, SD=1.08), and resource finding skills (mean=3.30, SD=1.06).

If the THE were to be implemented again, teachers (Obs=10) would focus on teaching grammar (mean rank=2.90) and writing (mean rank=3.00), followed by vocabulary and reading (mean rank=3.50). They would assign more homework related to writing and provide more feedback on writing (72.7%). About half of the teacher respondents would also like to apply stricter marking criteria for THE.

Qualitative results: interviews and open-ended questions in surveys

The 11 student participants were coded *S1*, *S2*. . . *S11*, with *S1* representing student 1, and so forth. Nine out of 11 students study Japanese, one French, and one Spanish. The five teacher participants were coded *T1*, *T2*. . . *T5*. Three out of five teach Mandarin, one French, and one Japanese.

Impact of THE on learning and teaching

Two students reported a positive impact on their learning attitudes because THE gave them less pressure (*S1*) and therefore made them “love their second language” (*S2*). In

contrast, the majority of student participants reported negative attitudes toward language learning when THE was implemented. They felt less motivated (S3, S6, S9) and spent less time on learning (S3, S5, S8); they worked much harder and spent more time on remembering vocabulary for ICEs (S4, from both survey and interview responses). However, with THE, students' learning changed to focus more on expressing themselves in the THE writing tasks (S6).

Most student interviewees did not change their learning practices even though since the beginning of the semester they had been informed that the exam would be a THE, because they felt they needed to "learn grammar, vocabulary to compose [their] own paragraph" anyway (S4); and because they were learning to develop the ability to communicate (rather than to merely pass exams) and the class contact hours were the same (regardless of exam format) (S2). They had interest and motivation (S3, S4, S5, S7): "learning language needs long-term effort and daily accumulation of knowledge" (S8).

The main change in learning practice happened during exam preparation. Most of the interviewees said that they felt less stressed because they did not need to remember words with a 24-h THE (S2) and could read the textbook during the exam (S3). Some of them did less preparation (S4) or even stopped reviewing (S9).

As for teaching, most teacher interviewees did not change their teaching methods because "[they] teach according to the learning outcomes;" that is, they teach the four skills (listening, reading, speaking, and writing) to prepare students for communication rather than for exams (T1). They did not want students to focus on exams too much (T4).

However, one teacher changed teaching to focus on oral and writing (T3). Others gave students more writing tasks or reading practice opportunities and tasks to develop their vocabulary during the teaching weeks to help them cope with the expected changes in assessment (T1, T2). No matter what teachers did during the semester, they focused on preparing the students for the exam when it was closer to the exam period (T1, T2, T3, T5).

Exam results

One student participant confirmed that she obtained higher marks in THE than ICE because she was better at using the language for communication rather than answering detailed grammar questions (S1). Nonetheless, most students believed that they were disadvantaged in THE, because there was no reading and grammar questions to help them obtain a higher mark (S3) and their peers obtained higher marks with "perfect works" in THE (S4), which signals academic integrity concerns with regard to THE.

When talking about overall exam performance, students had blurred or even contradictory impressions about which format of assessment led to better results. One student (S1) mentioned that most of their classmates did not like THE because it only had writing tasks and they feared that they would not obtain a high mark. However, S1 clarified later that the overall students' performance was better in THE because some students obtained good results which they did not deserve. Other students had similar opinions: on the one hand, they thought that the reading and grammar tasks in ICE help students obtain higher marks as these questions are objective and one could get full marks if the response is correct. On the other hand, they also thought that THE "enabled" weaker

students to ask for external help because there was enough time to do so and there was no supervision, whereby weaker students could also get good marks, causing an increase in average marks under THE compared with ICE (S5).

Teacher T3 agreed that the reading task in ICE did help a certain group of students who were in the second class (50–69 out of 100 marks) to obtain a higher mark, while cheating helped lower-achieving students to get higher marks.

Other teachers (T1, T5, T6) reported that, according to their own impression, there were no significant differences between students' exam results in THE and in ICEs. However, one of them mentioned that she heard students from other languages performed much better in the THE than in ICEs (T1).

Exam design and fairness

Students had polarized attitudes toward THE. Some were enthusiastic about THE, because it made the exam easier. Others hated THE and questioned its validity because they believed that it was not a “serious exam” (S4).

Most student interviewees preferred traditional ICE because “there’s less chance of cheating,” “it’s efficient,” and it has stood the test of time (S3).

For THE design, many students wanted the reading task. This is not only because reading skills are an important component of language learning (S2), but also because examiners' evaluation of the answers is standardized and “objective” and students have a chance at obtaining full marks (S3).

In the survey, some students expressed preference for different question types for different exam formats: THE should have writing tasks, as other tasks such as reading are “inefficient” (survey result), while for ICEs, they think that it is better to test “grammar and the knowledge points from the textbook” and it can be a fair method to test students' language level (S8).

Two students had a “bigger picture” about exam design. One student (S7) mentioned that “language is a tool for communication” so regardless of exam format, a language exam should include tasks that involve situations “a person might meet in the world of work.” Another student (S8) recommended having both THE and ICE because they required different skills from students and these skills are all useful.

Teachers preferred traditional ICE to THE, and thought that the latter was only for emergency use. They thought that for THE, duration should be shortened or more tasks should be given, but that the use of writing tasks per se was appropriate (T3). Writing based on a stimulus text not only tests writing skills, but students also need to understand the main points of the stimulus text and respond to it, so it also tests reading skills, analytical skills, and other higher cognitive skills like critical thinking skills (T3).

All teachers were satisfied with the THE design with writing tasks. One teacher mentioned that the reading stimulus text could be longer to integrate more reading skills assessment into the exam. Some teachers also recommended that ICEs have the same design as THE because the writing tasks in THE also evaluate grammar and reading in a more communicative way (T3, survey result).

To tackle the cheating issue in THE, one teacher recommended talking with the students about cheating and its consequences (T3). Another teacher (T4) suggested looking

into the overall design of the assessment. According to him, having the (closed-book) oral exam, which is worth 50% of the overall mark, is important because oral exam is “the most challenging” for students and they would “try to study more [for] oral exams” (T4).

Discussion

Impact of THE on learning and teaching

As presented in “[Impact of THE on learning and teaching](#)”, about 45% of the student respondents self-reported that, under THE, their learning foci were different from those under ICE. Specifically, in 2021–2022, when ICE was reinstated, they reported spending significantly more time on reading practice than in 2020–2021. The interview results in “[Impact of THE on learning and teaching](#)” provided a possible explanation for this difference. In 2020–2021, knowing in advance that THE would be implemented might have impacted some students’ learning attitudes and practices, because, for example, they knew they would have time to consult learning materials during the exam (see also Agarwal & Roediger, 2011; Durning et al., 2016). When ICE was reinstated in 2021–2022, students may have realized that THE had only been implemented as a temporary measure and thought it would be wise to study more than they had with THE in 2020–2021 to maintain a good level of performance. Moreover, the increase in reading practice time also mirrors the change in exam format: reading tasks appeared to be absent in the 2020–2021 THE but then returned as a significant part of the ICE in 2021–2022.

Nevertheless, albeit seemingly missing from the 2020–2021 THE, reading was in fact still a crucial part of the assessment, because students needed to read and comprehend the stimulus text in order to write an appropriate response. Some students did not seem to understand this point. Some student interviewees (S3, S4) frequently mentioned that the THE did not include a reading task, which means that they did not fully understand the role of the reading stimulus in the exam: their reading skills were still being assessed despite the lack of traditional tasks of reading comprehension, such as multiple-choice, true-or-false, and fill-in-the-blank questions. Interestingly, there is evidence that our THE task format had some success in assessing reading skills: under THE implementation, a teacher participant (T3) felt students’ reading skills were not well developed and that many students did not fully understand the reading stimulus and in some cases went off topic or failed to respond to it in full. While it is unclear whether there was a connection between students’ insufficient understanding of how reading was being assessed in the THE and their performance in that aspect, the current observations revealed that students’ understanding of an exam format may still be limited even with access to all relevant information; sometimes, what is obvious to teachers might not be as obvious to students. Consequently, it may be beneficial to provide students with more explicit explanations on certain aspects of an exam, so that students fully understand its requirements and expectations (see also Durning et al., 2016).

Furthermore, as Biggs and Tang (2011) pointed out, teachers should see the intended learning outcomes as the key element of their teaching (p. 197). Our teacher participants reported that they did not change their teaching methods, out of the belief that teaching should not be exam oriented, but be learning outcome oriented. However, students may think otherwise: they “learn what they think they will be tested on” (Biggs & Tang, 2011,

p. 197). Our student participants did exactly that. When ICE (along with the reading comprehension tasks) was reinstated, students spent significantly longer time on reading skills in self-study. Agarwal and Roediger (2011) compared students' learning habits and exam performance when students had different expectations for the final exam (closed-book vs. open-book) and found that "students' study habits may be based, in large part, on the perceived difficulty of a final test" (p. 850) and that the majority of students who were not informed about a specific type of final exam expected a closed-book exam (p. 849). As a result, the authors recommend that "teachers give closed-book tests or at least do not announce in advance that they will be giving open-book tests" (p. 850). However, in the UK higher education system, including at the university where the current study was conducted, students are to be informed of the type of exams they will undertake at the beginning of the module. To compensate for any potential impact of THE on learning, Agarwal and Roediger (2011) suggested that "teachers administer frequent quizzes" to improve long-term retention (p. 850).

Exam results

The comparison between THE and ICE exam results in "[Exam results](#)" showed that there was an apparent drop in the average marks for French and Spanish learners after the reinstatement of ICE in 2021–2022, while Japanese learners' average marks increased. The average mark could have been influenced by many factors such as the difficulty of the papers, which was not taken into consideration in this research. Therefore, we will limit our discussion to the analyses of the exam rankings.

According to the ranking analysis in "[Exam results](#)", the Spearman correlations for THE were much lower than those for ICEs, an indication that students' rankings in the ICEs were more stable compared with those in the THE. The unstable results in THE could be related to many factors which need further investigation but, importantly, it indicates that this type of THE is not as stable as ICE as an assessment tool for students' language skills at our institution. This result of our study reminds us to be cautious about the notion of using THE as the sole evaluation method for language modules.

Students' written THE and ICE rankings were weakly correlated for French and Spanish, but correlation was higher for Japanese. For French and Spanish, the results show that students' performance was quite different under different written exam formats. These differences mirror the interviews with students and teachers in "[Exam results](#)". Most interviewees believed that there were differences in exam performance between THE and ICEs. According to their belief, the types of exam tasks and dishonest behaviors were the two main factors that could have influenced the exam results. Based on their opinions, on the one hand, the lack of reading tasks in the THE lowered the mark of the top- and middle-achieving student groups. Karagiannopoulou and Milienos (2013) also found that THE benefited different student groups differently, depending on their approach to learning and their preference of exam format. We should thus take student's individual differences into account in assessment procedures as suggested by Myyry and Joutsenvirta (2015). On the other hand, the suspected dishonest behaviors were believed to have increased lower achievers' marks. One fact that could be considered consistent with this belief is that the top- and middle-achievers' results did not

change as dramatically as those of the lower achievers. The dishonesty issue will be further discussed in “[Exam design and fairness](#)”.

Curiously, our Japanese language students’ performance in THE was strongly correlated with their performance in ICEs but our current data do not seem to effectively explain this difference.

Overall, students’ results in the closed-book oral exams showed strong correlations across all semesters. When comparing students’ exam results in 2020–2021 and 2021–2022, the correlations were stronger when both written and oral results were combined in the analysis (i.e., when the final overall results of a module were used), compared with when only the written exam was considered. Crucially, this means that the oral exam (which is a closed-book in-class exam and is conceptualized as always being an ICE; see end of “[The implementation of THE](#)”) could mitigate the THE instability issue discussed earlier in this section. This result constitutes a significant complement to the conclusion of Durning et al. (2016) and Johanns et al. (2017) that a combined approach (of THE/OBE or ICE/CBE) could be more effective in assessing different competencies.

Exam design and fairness

According to the survey results shown in “[Exam design and fairness](#)” and the interview results in “[Exam design and fairness](#)”, teachers and most students preferred ICE in general as they believed it is much fairer. Fairness was the main concern students reported, and they felt that ICE is fairer because it has been tested over a long period of time; Gamage et al. (2020) also stated that ICE is a more secure exam method in the sense that academic misconduct is less likely to happen in an invigilated environment.

Many studies have recorded self-reported cases of some form of cheating. As reported by the International Centre for Academic Integrity (ICAI, n.d.), more than two-thirds of college students have self-reported cheating behaviors and cheating is still on the rise. Khan and colleagues (2022) commented that the shift to remote learning “comes with its own challenges, particularly in academic integrity during assessments, like the issue of academic dishonesty” (pp. 18–19). To tackle this, the language center tried various methods as explained in “[The implementation of THE](#)”. Some teachers also talked with students about cheating and its consequences, like suggested by McCabe et al. (2012) and by participants in Erguvan (2022). Despite these efforts, however, misbehavior in THE was still the biggest concern of our participants. Certainly, in the post-COVID digital age, more research and more experimentation of different strategies to avoid academic misconduct will be necessary. In these endeavors, just like in the current study, student voices should continue to be considered and their involvement should be encouraged (Azizi, 2022).

As far as second language learning is concerned, most of the participants of this study thought that ICE is not only fairer but more appropriate than THE. This opinion might be due to two reasons:

- 1) The nature of second language learning and testing. The key learning outcome tested in the language exams in the current study is the ability to express oneself, and especially at beginner and elementary levels, topics are related to everyday life. It is easy for a student taking a THE to obtain help from a native or proficient user of the tar-

get language and this would not be detected by any anti-cheating software. This practice is known as “contract cheating” (Ahsan et al., 2021, p. 523). Conversely, contract cheating can be more difficult in THE for other disciplines as they require specialized knowledge and, sometimes, citation of examples and references (Gamage et al., 2020);

- 2) Second language proficiency necessitates gradual learning. Learning a language is not a one-semester or 1-year effort. It requires time and students need to demonstrate progress from one level to the next by demonstrating that they have acquired the required skills. If they do not have the necessary foundations, the following level of study will be more challenging. Enhancing the base knowledge is the key and ICEs are thought to be associated with a greater amount of study and produce better learning in a university-level environment (Marsh, 1984).

The two reasons above might also apply in other disciplines, like science courses where there is clear progression in knowledge and skills and whose exams could require mostly calculations and solution of problems without the need for extensive reading and citation. ICEs might work better than THE for evaluating students’ learning in sciences, but more research is needed both for language and scientific disciplines.

Nevertheless, during COVID-19, at the language center, THE was thought to be the only possible solution in place of ICE. To maximize THE validity and reliability, the language center relied on changing the exam design and duration. As Cleophas and colleagues (2021) mentioned, a key to avoiding fraud in the first place is a suitable design of online exams. Teacher participants in this study agreed that THE should test productive skills such as writing, and that it would be better to also test student’s receptive skills such as reading. Our THE writing task with a reading stimulus required students to understand the main points of the stimulus and write a response based on their own experience or express their own opinions. It incorporated reading skills into the exam and was thus evaluated internally as an appropriate design. Teacher participants felt that this kind of task enabled them to assess language skills, as well as such key skills for university students as critical thinking and information management skills (see also López et al., 2011). They also thought that this type of task enhanced deep learning as students engaged with higher-order skills like analytical skills, agreeing with the conclusion of Johanns et al. (2017).

Regarding exam duration, over half of students and most teachers who responded to the survey agreed that 24 h was too long for the THE, because it gave time for dishonest behaviors to take place. Respondents suggested 2 to 4 h for THE, more specifically about 2 h to answer the exam questions and some extra time to deal with any operational tasks, such as uploading the answers to the exam platform. These responses are supported by findings of some existing studies (Ng, 2020; Spiegel & Nivette, 2021; Tam, 2022), which also recommended setting tight time restrictions for THE.

Besides the written exam design, it is also important to look into the overall module assessment design. One teacher mentioned that when implementing THE, it is essential to have another exam component, in our case a closed-book in-class oral exam, to ensure the reliability and validity of the overall module assessment. As the participant explained, students will study more if they need to take an in-class oral exam which

requires them to memorize and use vocabulary, grammar, and phrase expressions, skills they would use also in the THE. Therefore, having to take the THE with an oral exam, students are more likely to study the language rather than merely rely on cheating. This result is echoed by the conclusion of Durning et al. (2016) that the combined use of OBE and CBE could be effective as “OBEs and CEBs can contribute to an assessment program in part because of their complementary pros and cons” (p. 588), and by Johanns and colleagues (2017) who favored the use of a mixed method of examinations throughout the course of a nursing program.

Limitations

The biggest concern of our research participants is exam fairness and suspected academic dishonesty. However, it was difficult to obtain data of cheating behavior in THE and we were unable to fathom the real impact of academic misconduct on THE results. While this study provides insights into the impact of THE in second language learning and teaching, there are important limitations to consider that might influence the interpretation of our findings. First, the interviewees study and teach different languages, and because of the limited number of participants from each language, we could not evaluate whether their opinions and experiences were related to specific characteristics of the teaching and/or learning of a certain language, or whether they were shared by more students or teachers at the institution. The limited sample size of interviewees and survey respondents also means that their views and experiences may not fully represent those of all our language students and teachers. Second, the self-reported nature of the survey and interview data may have introduced bias and led to inaccuracies. Third, we did not investigate the impact of factors such as students’ well-being (Stowell & Bennett, 2010), motivation, or learning style (Spiegel & Nivette, 2021). Although we used exam result data from the same cohort of students, these data span two academic years. This study did not consider any changes that individual students and/or the cohort may have undergone during this time. Fourth, the teacher participants come from diverse cultural backgrounds, which might have influenced their teaching and marking (Bianco & Crozet, 2003). Fifth, the study was conducted on a transnational campus, which has its own features that may not be found in other higher education contexts, rendering it necessary to take extra caution against overgeneralization of our findings. Finally, the study was conducted in the unique context of the COVID-19 pandemic. Our results about THE’s impact on teaching and learning cannot be generalized to non-pandemic situations.

Conclusions

This study investigated the impact of 24-h take-home exam in the field of language education on the China campus of a British university by analyzing students’ exam performance, teachers’ teaching methods, and students’ learning strategies under THE implementation. Our findings show that the implementation of THE during the pandemic did not change teachers’ teaching foci but expected exam format influenced many students’ learning practices: students tended to spend more time on skills that they anticipated would be tested in exam. Also, students and teachers believed that cheating was a major issue under THE.

The current study also found that students' rankings exhibited fairly strong stability in ICEs, but such stability was not found with THE. When results of THE and ICE (oral) components were combined, however, the overall ranking stability greatly improved, suggesting that the oral ICEs mitigated the ranking instability associated with the written THE in our study.

Student and teacher participants preferred ICEs for evaluating learners' language levels, but teachers still considered the THE with two writing tasks based on reading stimuli an appropriate tool to evaluate language learning as they involve higher-level thinking skills. In addition, some students and teachers favored the design of language assessment with a combined use of ICE and THE, based on the consideration that when THE was implemented, the in-class closed-book oral exam could enhance the overall validity of the assessment for a language module. Strategies like this could potentially improve the usability of THE as a formal assessment instrument, and future research could explore the effectiveness of various strategies used in operations such as exam design, administration, and grading to offset the shortcomings of THE.

Abbreviations

AY	Academic year
CBE(s)	Close-book exam(s)
ICE(s)	In-class exam(s)
OBE(s)	Open-book exam(s)
THE(s)	Take-home exam(s)

Acknowledgements

We would like to express our deepest gratitude to Dr. Giovanna Comerio for her encouragement and invaluable feedback throughout the entire study.

Authors' contributions

XY acted as the project administrator and made substantial contributions to the formation of research aims, the study design, the acquisition, analysis and interpretation of the exam results, interview, and survey data, and to manuscript writing. YY contributed to the development of research aims, the quantitative analysis and interpretation of exam results, and to manuscript writing, review, and revision. YQ contributed to the formation of research aims, the study design, and qualitative analysis of the interview data. ZH contributed to survey and exam data analysis and presentation of quantitative analysis results. All authors read and approved the final manuscript.

Funding

None.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Ethics approval was obtained from the UNNC Research Ethics Subcommittee.

Consent for publication

Consent to publish has been obtained from the participants.

Competing interests

The authors declare no competing interests.

Received: 2 February 2024 Accepted: 12 June 2024

Published online: 26 June 2024

References

Agarwal, P., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, 19(8), 836–852. <https://doi.org/10.1080/09658211.2011.613840>

- Ahsan, K., Akbar, S., & Kam, B. H. (2021). Contract cheating in higher education: A systematic literature review and future research agenda. *Assessment & Evaluation in Higher Education*, 47(4), 523–539. <https://doi.org/10.1080/02602938.2021.1931660>
- Azizi, Z. (2022). Fairness in assessment practices in online education: Iranian university english teachers' perceptions. *Language Testing in Asia*, 12(1), 1–14. <https://doi.org/10.1186/s40468-022-00164-7>
- Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society*, 16(2), 296–298. <https://doi.org/10.1111/j.2517-6161.1954.tb00174.x>
- Bengtsson, L. (2019). Take-home exams in higher education: A systematic review. *Education Sciences*, 9(4), 267. <https://doi.org/10.3390/educsci9040267>
- Bianco J., & Crozet C. (2003). Teaching invisible culture: Classroom practice and theory. Language Australia Ltd.
- Biggs, J., & Tang C. (2011). Teaching for quality learning at university. McGraw-Hill Education.
- Bone, A., & Maharg, P. (2019). *Critical perspectives on the scholarship of assessment and learning in law* (Volume 1). England: ANU Press. <https://doi.org/10.22459/CP01.2019>
- Braselmann, S., Mathieson, J., & Moisch, O. (2022). Multimodal take-home exams in online teaching and beyond: Constructive and professional alignment in teacher education. *Zeitschrift für Hochschulentwicklung*, 17(1), 87–102. <https://doi.org/10.3217/zfhe-17-01/06>
- Chan, C. K. Y. (2022). A review of the changes in higher education assessment and grading policy during Covid-19. *Assessment & Evaluation in Higher Education*, 48(6), 874–887. <https://doi.org/10.1080/02602938.2022.2140780>
- Clark, T. M., Callam, C. S., Paul, N. M., Stoltzfus, M. W., & Turner, D. A. (2020). Testing in the time of COVID-19: A sudden transition to unproctored online exams. *Journal of Chemical Education*, 97(9), 3413–3417. <https://doi.org/10.1021/acs.jchemed.0c00546>
- Cleophas, C., Hönnige, C., Meisel, F., & Meyer, P. (2021). Who's cheating? mining patterns of collusion from text and events in online exams. *Informations Transactions on Education*, 23(2), 84–94. <https://doi.org/10.1287/ited.2021.0260>
- Croasmun, J. T., & Ostrom, L. (2011). Using Likert-type scales in the social sciences. *Journal of Adult Education*, 40(1), 19–22. <https://www.proquest.com/scholarly-journals/using-likert-type-scales-social-sciences/docview/1018567864/se-2>
- Csató, L. (2013). Ranking by pairwise comparisons for Swiss-system tournaments. *Central European Journal of Operations Research*, 21, 783–803. <https://doi.org/10.1007/s10100-012-0261-8>
- Dabestani, R., Taghavi, A., & Saljoughian, M. (2014). The relationship between total quality management critical success factors and knowledge sharing in a service industry. *Management and Labour Studies*, 39(1), 81–101. <https://doi.org/10.1177/0258042X14535160>
- Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. W. (2016). Comparing open-book and closed-book examinations. *Academic Medicine*, 91(4), 583–599. <https://doi.org/10.1097/acm.0000000000000977>
- Erguvan, I. D. (2022). University students' understanding of contract cheating: A qualitative case study in Kuwait. *Language Testing in Asia*, 12(56), 1–19. <https://doi.org/10.1186/s40468-022-00208-y>
- Fernald, P. S., & Webster, S. (1991). The merits of the take-home, closed-book exam. *The Journal of Humanistic Education and Development*, 29(4), 130–142. <https://doi.org/10.1002/j.2164-4683.1991.tb00017.x>
- Gamage, K. A., Silva, E. K. D., & Gunawardhana, N. (2020). Online delivery and assessment during COVID-19: Safeguarding academic integrity. *Education Sciences*, 10(11), 301. <https://doi.org/10.3390/educsci10110301>
- Gao, X. (2006). Understanding changes in chinese students' uses of learning strategies in China and Britain: A socio-cultural re-interpretation. *System*, 34(1), 55–67. <https://doi.org/10.1016/j.system.2005.04.003>
- Gliem, J. A., & Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Midwest Research-To-Practice Conference in Adult, Continuing, and Community Education*, 1, 82–87.
- Gunawan, T. J., Wang, J., & Liao, P. C. (2022). Factors of project-based teaching that enhance learning interest: Evidence from construction contract management course. *Sustainability*, 14(22), 15314. <https://doi.org/10.3390/su142215314>
- Harrison, D. (2020, April 28). Online education and authentic assessment. *Inside Higher Education*. Retrieved December 13, 2022, from <https://www.insidehighered.com/advice/2020/04/29/how-discourage-student-cheating-online-exams-opinion>
- International Centre for Academic Integrity (ICAI). (n.d.). *Facts and statistics*. Retrieved December 22, 2022, from <https://academicintegrity.org/resources/facts-and-statistics>
- Jacobs, A. D. (2021). Utilizing take-home examinations in upper-level analytical lecture courses in the wake of the COVID-19 pandemic. *Journal of Chemical Education*, 98(2), 689–693. <https://doi.org/10.1021/acs.jchemed.0c00768>
- Johanns, B., Dinkens, A., & Moore, J. (2017). A systematic review comparing open-book and closed-book examinations: Evaluating effects on development of critical thinking skills. *Nurse Education in Practice*, 27, 89–94. <https://doi.org/10.1016/j.nepr.2017.08.018>
- Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39(1), 31–36. <https://doi.org/10.1007/BF02291575>
- Karagiannopoulou, E., & Milienos, F. S. (2013). Exploring the relationship between experienced students' preference for open- and closed-book examinations, approaches to learning and achievement. *Educational Research and Evaluation*, 19(4), 271–296. <https://doi.org/10.1080/13803611.2013.765691>
- Khan, Z. R., Priya, J., & Tuffnell, C. (2022). Culture of integrity – Institutional response to integrity during COVID19. *International Journal for Educational Integrity*, 18, 27. <https://doi.org/10.1007/s40979-022-00118-9>
- Krzywinski, M., & Altman, N. (2014). Points of significance: Nonparametric tests. *Nature Methods*, 11(5), 467–468. <https://doi.org/10.1038/nmeth.2937>
- Kvale, S. (2008). Doing interviews. *Sage Publications Ltd*. <https://doi.org/10.4135/9781849208963>
- Li, C., & Ruan, Z. (2015). Changes in beliefs about language learning among Chinese EAP learners in an EMI context in Mainland China: A socio-cultural perspective. *System*, 55, 43–52. <https://doi.org/10.1016/j.system.2015.08.010>
- López, D., Cruz, J., Sánchez, F., & Fernandez, A. (2011). A take-home exam to assess professional skills. *2011 Frontiers in Education Conference (FIE)*, F1C-1–F1C-6. <https://doi.org/10.1109/fie.2011.6142797>
- Marsh, R. (1984). A comparison of take-home versus in-class exams. *The Journal of Educational Research*, 78(2), 111–113. <http://www.jstor.org/stable/27540103>

- McCabe, D. L., Butterfield, K. D., & Treviño, L. K. (2012). *Cheating in college: Why students do it and what educators can do about it*. The Johns Hopkins University Press.
- McNamara, T. (2000). *Language testing*. Oxford University Press.
- Myrsky, L., & Joutsenvirta, T. (2015). Open-book, open-web online examinations: Developing examination practices to support university students' learning and self-efficacy. *Active Learning in Higher Education*, 16(2), 119–132. <https://doi.org/10.1177/1469787415574053>
- Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches*. Pearson.
- Ng, C. (2020). Evaluation of academic integrity of online open book assessments implemented in an undergraduate medical radiation science course during COVID-19 pandemic. *Journal of Medical Imaging and Radiation Sciences*, 51(4), 610–616. <https://doi.org/10.1016/j.jmir.2020.09.009>
- Quality Assurance Agency for Higher Education. (2013). *Review of UK trilateral education in China 2012: Overview*. Retrieved December 8, 2022, from [https://www.qaa.ac.uk/docs/qaa/international/tne-china-overview-\(1\).pdf?sfvrsn=e43ff481_2](https://www.qaa.ac.uk/docs/qaa/international/tne-china-overview-(1).pdf?sfvrsn=e43ff481_2)
- Quality Assurance Agency for Higher Education. (2018). *UK quality code for higher education: Advice and guidance*. Retrieved December 13, 2022, from http://www.qaa.ac.uk/docs/qaa/quality-code/advice-and-guidance-assessment.pdf?sfvrsn=ca29c181_4
- Raje, S., & Stitzel, S. E. (2020). Strategies for effective assessments while ensuring academic integrity in general chemistry courses during COVID-19. *Journal of Chemical Education*, 97(9), 3436–3440. <https://doi.org/10.1021/acs.jchemed.0c00797>
- Rich, J. D. (2011). An experimental study of differences in study habits and long-term retention rates between take-home and in-class examinations. *International Journal of University Teaching and Faculty Development*, 2(2), 121–129. Retrieved from <https://www.proquest.com/scholarly-journals/experimental-study-differences-habits-long-term/docview/1722618786/se-2>
- Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Sage.
- Şenel, S., & Şenel, H. C. (2021). Use of take-home exams for remote assessment: A case study. *Journal of Educational Technology and Online Learning*, 4(2), 236–255. <https://doi.org/10.31681/jetol.912965>
- Shin, J., & Park, Y. (2009). On the creation and evaluation of E-business model variants: The case of auction. *Industrial Marketing Management*, 38(3), 324–337. <https://doi.org/10.1016/j.indmarman.2007.06.017>
- Skovlund, E., & Fenstad, G. U. (2001). Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? *Journal of Clinical Epidemiology*, 54(1), 86–92. [https://doi.org/10.1016/s0895-4356\(00\)00264-x](https://doi.org/10.1016/s0895-4356(00)00264-x)
- Spearman, C. (2010). The proof and measurement of association between two things. *The International Journal of Epidemiology*, 39(5), 1137–1150. <https://doi.org/10.1093/ije/dyq191>
- Spiegel, T., & Nivette, A. (2021). The relative impact of in-class closed-book versus take-home open-book examination type on academic performance, student knowledge retention and wellbeing. *Assessment & Evaluation in Higher Education*, 48(1), 27–40. <https://doi.org/10.1080/02602938.2021.2016607>
- Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Journal of Educational Computing Research*, 42(2), 161–171. <https://doi.org/10.2190/ec.42.2.b>
- Tam, A. C. F. (2022). Students' perceptions of and learning practices in online timed take-home examinations during Covid-19. *Assessment & Evaluation in Higher Education*, 47(3), 477–492. <https://doi.org/10.1080/02602938.2021.1928599>
- Tao, J., & Li, Z. (2012). A case study on computerized take-home testing: Benefits and pitfalls. *International Journal of Technology in Teaching and Learning*, 8(1), 33–43. https://sictet.org/main/wp-content/uploads/2016/11/ijttl-12-01-3_Tao.pdf
- University of Nottingham Ningbo China. (n.d.). *About the university*. Retrieved January 23, 2023, from <https://www.nottingham.edu.cn/en/About/Who-we-are.aspx>
- Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure sciences*, 39(2), 163–173.
- Wadgave, U., & Kahairnar, M. R. (2019). Parametric test for non-normally distributed continuous data: For and against. *Electronic Physician*, 11(2), 7468–7470. <https://doi.org/10.19082/7468>
- Wellington, J. (2015). *Educational research contemporary issues and practical approaches*. Bloomsbury.
- Williams, J. B., & Wong, A. (2009). The efficacy of final examinations: A comparative study of closed-book, invigilated exams and open-book, open-web exams. *British Journal of Educational Technology*, 40(2), 227–236. <https://doi.org/10.1111/j.1467-8535.2008.00929.x>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.