SN Computer Science (2022) 3:270
https://doi.org/10.1007/s42979-022-01132-x

ORIGINAL RESEARCH

Exploring the Potentials of Performance-Centred Instructional Design in Online and Blended Learning: Students' Perspectives

Terumi Miyazoe[1]

Received: 27 December 2021 / Accepted: 3 April 2022 / Published online: 5 May 2022

Abstract

This study aims to clarify the potentials of performance-centred instructional design in online and blended learning. It asserts its contribution in that no research so far available has explicitly investigated how students—'evaluatees'—evaluate the implemented course design with a 100% performance-based assessment (PA). The study consists of two parts: a comprehensive literature review on PAs, followed by a report on a survey to the students on performance-centred course experiences. The research was held in an undergraduate English programme at a Tokyo university. No test was conducted, and only performance-based evaluative methods were used throughout the two-semester course of one academic year. A survey was administered to the students after the course experience, obtaining 67 valid responses. AI research tools were also applied to the analyzes, to explore their future use. A high level of positivity toward the PA course design was obtained. Besides, the design succeeded in building students' self-efficacy and helping them become more strategic in using the language. The perceived progress was also confirmed by an objective test held outside the current research. Furthermore, the students admitted some utility in test-based assessment, proving, on average, that the combinatory design of 75–85% PAs and 15–25% tests would be ideal. PA-centred course design has significant potential to deepen students' learning. It provides an antithesis to a heavily test-centred teaching approach, which could limit students' learning. It suggests that digital-based PAs be viable solutions, when meeting for tests is not feasible under emergencies, including worldwide pandemics.

Keywords Performance assessment · Blended learning · Learning outcomes · Instructional design · AI · Emergency

Introduction

This study is motivated by the desire to find a potentially improved alternative to the test-centred teaching approach, using performance-based assessment (PA) in higher education. The necessity of doing this research comes from our socio-cultural background: the Japanese society is known for its strong emphasis on winning competitive, high-stakes examinations [1, 2], a social norm historically linked to a traditional imperial examination system to become a high-ranking civil servant. This trend is also shared by neighboring Asian countries [3, 4], and test-takers and the endorsement of the authority of tests and test-makers (teachers) carry significant social value. Within this socio-cultural climate, the current study explores the potentials of the PAs in teaching and learning compared to the test-centred teaching approach in higher education and skill-based blended courses.

This paper has been elaborated under the unique pandemic situation. The core data had been obtained before the start of the COVID pandemic in 2019. After a 1-year interruption due to the chaotic state of education in 2020, the author took the work up amid the recovery, aiming for the new normal in 2021. Therefore, the data in this study bear a special significance for its preservation of pure data, before the emergency. The study can be understood as a screenshot of a cultural environment that was acutely altered by the pandemic.

Literature Review

Definition of a Performance-Based Assessment (PA)

In this study, performance-based assessment (PA) refers to a concept in which grading is decided by accumulating

✉ Terumi Miyazoe
t.miyazoe@rs.tus.ac.jp

[1] Tokyo University of Science, Tokyo, Japan

students' performance and products as an alternative to the periodic traditional paper-and-pen written examinations that rely more on rote memory. In a chapter devoted to PA, Cantu and Warren [5] review and list commonly accepted definitions of PA, among which the following is presented as the definition most closely matches the aim of the current research:

> Performance assessment, also known as alternative or authentic assessment, is a form of testing that requires students to perform a task rather than select an answer from a ready-made list. For example, a student may be asked to explain historical events, generate scientific hypotheses, solve math problems, converse in a foreign language, or conduct research on an assigned topic [6].

Some may consider that PA was recently introduced as an antithesis to objective testing difficulties [7, 8]; however, PA precedes these norm-referenced tests. A comprehensive work on the history of PA by Madais and O'Dwyer [9] traces its origin to 210 BCE, during the Han Dynasty in China, where the assessment was designed to select competent civil and military servants. The author details the use of this method for assessment up to the late twentieth century via a summary of detailed historical records: one critical point they reveal is that during the Sung Dynasty (960 CE–1279 CE) in China, questions demanding rote answers were replaced, since the scoring method to measure 'reasoning ability' could be highly subjective. In other words, chronologically, objective testing is an antithesis to PA: the current PA is, in this sense, some revival of our ancient evaluation methodology in 2000 years of human civilisation.

Prior Research on Performance-Based Assessment in Education Assessment

A comprehensive literature review was conducted with a particular focus on PA in language education, because this is the most relevant area of study experimented on in the current paper. Research on PA and criterion-reference tests is abundant in educational assessment and language education. To find the most relevant data, 12 journals that were available from Scopus [10] were selectively referenced [11]; the keywords 'performance assessment', 'objective test', and 'student' were used to find articles that explicitly focused on students' perspectives, per this study's research concerns. The searches were undertaken consecutively on a specific day in February 2020 to avoid any fluctuation in the coverage caused by the regular updates of the database.

Table 1 summarizes the search results for this study. The column 'Entry' refers to the total number of research articles as of the search date in each journal: the column 'Hit' presents the number of relevant articles for this study's literature review among them. All 17 articles explicitly or inexplicitly examined both PA and the norm-referenced approach in different ways, in that they do not regard measurements/assessments as self-evident, and they include students', i.e., examinees perspective. These 17 studies were thoroughly examined and categorized into three groups depending on the focus of their research questions or designs: overarching, measurement-focused, and evaluation process-focused.

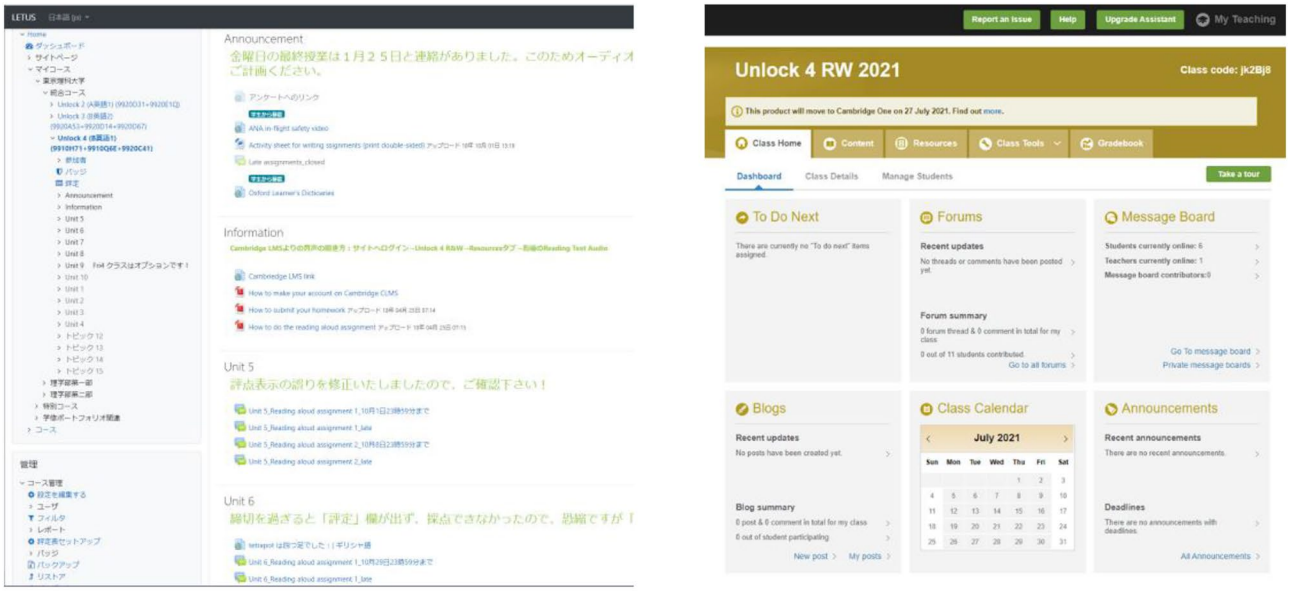
The first category comprised those providing overarching arguments regarding PA. In the US, Hambleton [12] presents a historical literature review to conclude a large volume of research on criterion-based measurement that was undertaken from the 1970s to 1984, which becomes the base of the widespread support for the coexistence of norm-referenced

Table 1 Literature review results from Scopus.com (as of 27 Feb, 2020)

	Journal	Entry	Hit	Country
Educational measurement and assessment journals	Assessment in Education: Principles, Policy and Practice	544	1	Canada
	Assessment and Evaluation in Higher Education	1416	0	–
	Educational Evaluation and Policy Analysis	734	1	US
	Educational Measurement: Issues and Practice	1071	8	All US
	Journal of Educational Measurement	1548	1	US
	Practical Assessment, Research and Evaluation	358	0	–
	Applied Psychological Measurement	1734	1	US
	Psychometrika	3082	0	
Language testing journals	Language Testing	777	2	US 1, Japan 1
	Language Assessment Quarterly	270	1	Japan
	Assessing Writing	408	2	US 1, Canada 1
	Language Testing in Asia	160	0	–
Total		12,102	17	

and criterion-based measurements. Camara and Brown [13], at the American Psychological Association (APA), examined the changing concepts of employment testing using PA, implying that PA research in education is lagging behind that in industry. Cizek [14] synthesized the 'standard-setting standards' (p.13) for assessment guidelines with an emphasis on the right match between the assessment method and the assessed, including PA and the norm-referenced mode. Nicholas and Sugrue [15], also in the US, examined the traditional test-making process to propose a 'higher fidelity' construct-centred approach to cognitively complex constructs, regardless of the use of PA or multiple-choice testing. Another US work by Hambleton et al. [16] examined 10 PA guidelines and their pros and cons, while Koretz [17], who promoted multiple measurements, observed that a single measurement use could contradict article 13.7 of the *Standards for Educational and Psychological Testing* [18] (article 13.9 in the most recently amended version [19]). Finally, via a comprehensive literature review from 1991 to 2012, Gotch and French [20] identified 36 teacher 'assessment literacy measures', hypothesizing that literacy and students' learning outcomes could be correlated, though this was supported by just one study Koh [21] in the review at that time.

In the second category of examining the validity and possible scale bias of PA, another work by Cizek [22] in the US investigated the standards for setting passing scores (pass/fail) for PAs by reviewing several methods/models, including norm-referenced. Klein et al. [23] examined the validity of PA and objective methods among gender and racial/ethnic groups. They found that the measurement type did not affect the differences among the groups (Figs. 1 and 2



Online course on the school Moodle

Online course on Cambridge LMS

Fig. 1 An image of the online part of the blended course

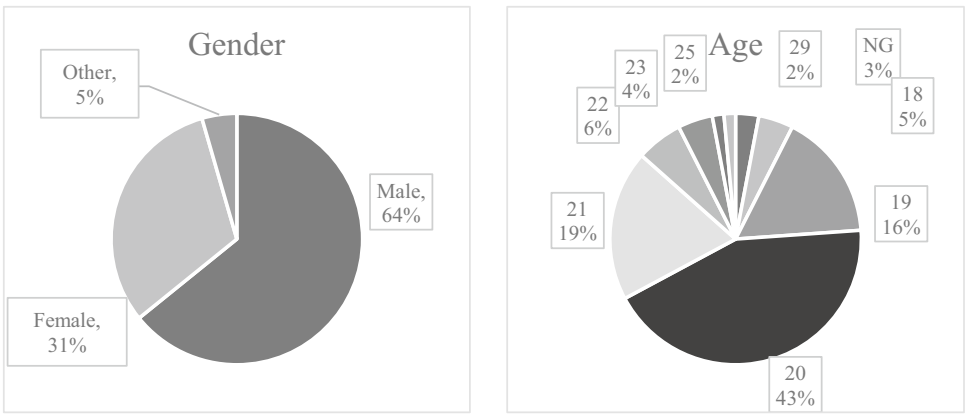


Fig. 2 Respondents' profile of students in the study

in that paper): traditionally, it is explained that females and minority ethnic groups have been believed to score lower in objective tests. In Canada, Fox and Cheng [24] examined the possible disadvantages that PA may present for English as a second language (L2) when it has been designed for English as a first language (L1) learners. Further, Barkaoui [25] explored the possible bias residing in the holistic and multiple-trait rating of PAs for English as a foreign language (EFL) essay writing, while, in Japan, Kozaki [26] attempted to improve cut estimates of PA scale formula workable for small-scale, low-cost use.

The third category comprised those works that addressed rater bias: the possible fluctuation in scoring by raters using the same PA scaling. In Hawaii, Kondo-Brown [27] investigated three raters for Japanese writing as a second language to observe whether there were individual rater characteristics of harshness or leniency for specific semantic or syntactic areas. Meanwhile, in Japan, Matsuno [28] compared PA evaluations by individuals, peers, and teachers in English writing and concluded that students tend to evaluate their self-performance lower than the other two groups. Penny and Johnson [29] probed the quality of PA with several factors, using the Monte Carlo computer simulation method to conclude that expertise in writing assessment would be the most critical factor to ensure the PA assessment. Lastly, Kane [30] found that the possible bias involved in handling borderline cases in criterion-based assessment falls within the category of rater bias because judges' decisions were concluded to have a considerable influence.

In summary, the arguments between PA and objective testing seem to be at the stage of acknowledging that both have pros and cons, to the extent of focusing more on specific issues regarding the inner mechanisms of scales/raters, as well as conditions in which they are used and interpreted. No research design has been found that has explicitly investigated how students—'evaluatees'—evaluate assessment implementation design and what learning outcomes they would have from this experience. Cizek [22] suggested that 'a potentially fruitful line of research lies in combining the needs of standard setters with the knowledge base in instructional design' (p. 28). The present research was conducted in this vein, in which PA was fully embedded in the blended instructional design to examine its potential effectiveness.

Methods

Research Questions

Under the above-reviewed location of the study, this research tries to answer the following questions:

1. How do students evaluate the 100% PA-centred instructional design?

2. How could we incorporate PA elements in future digital learning?

Research Tools

The present study employed two relatively new research tools/concepts—an online survey with a slider bar question format and AI text mining. The online survey system Survey Monkey (https://www.surveymonkey.com/) was used to provide sophisticated survey making, delivery, and analysis functions. An online survey also has several other merits in that it offers anonymity, accessibility, and cost and time savings [31]. The survey's slider bar question format asked respondents to mark their agreement level in integers from 0 to 100. Despite its popularity and wide use, there have been arguments about the validity of Likert-type scale questions [32, 33]; statistically, the slider bar offers a more precise and granular representation than categorical rating. The author tested online surveys and slider bar responses in previous research, confirming that their functionality would offset the weaknesses of the paper-based Likert-type survey method [34].

AI text mining was used to analyze the text comments to the survey's open-ended questions. A beta version of the AI-based text mining system (User Local: AI text mining: https://textmining.userlocal.jp/) that also processes Japanese language text was available. The system provides results ranging from a basic word cloud, word frequency, and concurrent keywords to an advanced negative/positive emotional analysis, hierarchical clustering, digest, and highlighting. Text analysis for qualitative data dates to its theorization as a grounded theory by Glaser [35]: the author of this paper practiced the manual coding technique using computer-based analytics such as ATLASti and SPSS Text Analytics [36]. AI text mining has recently been drawing much attention in Japan; however, its use remains limited to corporations and laboratory research requiring extensive funds and skills such as Visualization Engine (https://www.pa-consul.co.jp/) and KIBT (https://www.scsk.jp). The present study also tests its usefulness in small-scale research by individual researchers.

PA-Centered Blended Instructional Design

In this study, students of seven classes followed the same course design of a blended format. All the courses used the same core course textbooks (Unlock Reading and Writing Series 2–4 from Cambridge University Press, 2014) but differed in target English proficiency levels, including A2, B1, and B2 in the Common European Framework of Reference (CEFR) for languages scales [37]. All students were part of the Faculty of Science of a Tokyo university, majoring in Physics, Chemistry, or Mathematics.

The courses were a blend of 30 in-class meetings over 1 year, with homework assignments comprising three to five essay compositions (writing) on paper and Cambridge LMS (https://www.cambridgelms.org/), course references, and weekly oral audio file submissions (speaking) via the learning management system Moodle (https://moodle.org/), as per the new design. The length of each writing assignment differs depending on the level of the courses: on average 50–100 words for A2 whereas 300–400 words for B2; the length for each audio assignment also differs: on average one minute for A2 whereas three minutes for B2 as an indication. The amount of time for checking/giving feedback to each assignment is non-negligible: the teacher makes it a rule to work on these at a regular pace of one hour per day, for example, throughout the semester to provide as timely feedback as possible but not too overwhelming. Moodle also served as a repository for the students' portfolios during the course. Figure 1 below presents an example image of the online part of the blended courses in the current study. As the same course content has been offered (with publisher's revisions), the image on the right is marked 2021 when the screenshot was taken.

The final course evaluation was solely PA-based, with no test components included. All the writing submitted by the students was edited by the instructor and returned with numerical feedback: six points from the three-aspect evaluation method (syntax, content, and the goal of each unit) were used to attenuate subjectivity and add analytical features into the PA. A three-point feedback scale was used for speaking assignments graded roughly as satisfactory, neutral, and unsatisfactory; listening to each other's submission was recommended to the students for self-learning but was not compulsory. Additionally, although regular deadlines for assignment submission were set, late submission was accepted with a one-point deduction as a minor penalty for encouraging regular completion of tasks. The grading policy was explained clearly at the course orientation in all classes: the students could choose to take the current class or switch to another that was usually assessed via a high-stakes final examination as an alternative.

Analysis of PA Design in the Courses

The use of PA in the current study was analyzed, and the summaries are presented in Table 2. The analysis was made from the perspective of the two essential qualities, 'authentic task' and 'performer-friendly feedback', required to legitimately be a PA, as defined by Wiggins [38, p.21]. This process is critical because if the instructional design fails the PA criteria, this study will be less relevant to the field understanding: the assessment measurements seem to pass the necessary PA criteria, though there would be room for amendments in future practice.

Data Collection

For the PA-centered course design in this research to be evaluated, an author-designed survey was used to collect information from the students. The survey was administered during the last session before the final day of returning the course grades, thereby minimizing the possibility of the course grade influencing the evaluation of the course experience. To obtain a reasonable response rate, students were allowed to respond to the survey online or on paper [39]—the online survey system was identical to a PDF version of the questionnaire from the paper version. The survey contained 28 questions, and the system estimated that answering the survey online would require approximately six minutes. The survey asked for students' perceptions on specific aspects of (1) assignments (difficulty, content,

Table 2 Analysis of performance-based assessment in the course

	Authenticity	Feedback
Participation in-class activities (40%)	Task-based reading and listening activities situated in the context prepared in the course textbook Many occasions to answer using the target language in front of others	Self-correction by sharing the correct answers (if any) in-class activities A friendly atmosphere of the classroom to support and encourage participation
Writing composition (30%)	Topic provided from each unit's focused content Each student writes in their own words	0–6 points consisting of three aspects of contents, mechanics. and topical approach, plus written comments if necessary Detailed editing by the instructor Returned within one week after the due date
Audio submission (30%)	Read aloud, record, and submit narration after practicing with authentic narration provided by the course textbook Shadowing practice as a challenge option from the fall semester Audio sharing on the LMS	0–3 points plus written comments when necessary Returned within one week to 10 days after the due date

frequency, and method of completion), (2) self-evaluation based on 'can-do' notions regarding the changes in specific language areas, (3) evaluation on the appropriateness of PA in the course, (4) demographic features (age and gender), and (5) open-ended questions regarding the instructional design and management. An English translation of the survey can be obtained by contacting the author of this paper.

The course employed in this research excluded 'tests'. However, one among the seven classes was in a different division of the same Faculty of Science (in Table 3, the Applied Chemistry major class). In this particular class, other than the present study's course scheme, the students were required to pass the Test of English for International Communication (TOEIC) English proficiency test (https://www.ets.org/toeic) at the end of each semester. After the survey, the researcher obtained each student's signed written permission to use their TOEIC scores as a reference in calculating the class average to triangulate the survey results.

Results

Depending on the features highlighted in the discussions, results are presented as either figures or tables for higher comprehensibility.

Respondents' Profiles

The profile of the respondent students in the study is summarized in Table 3 and Fig. 2 below. Among the 67 valid responses, the male: female: no-wish-to-give student ratio was 64.2%:31.3%:4.5%, respectively, which approximates the students' gender profile similar to that published by the Faculty of Science (73%:27%) [40]. Among 67, two students chose no-wish-to-give choice ($N=65$, Min 18, Max 29, Mean 20.42, and SD 1.648): as a general profile, their ages vary from 18 to 29 years among which 83% were 19–21 years old. The students were taking English courses at different stages: the total number of respondents by level are presented in the Total column. They were divided evenly, with the CEFR Level A2 group slightly outnumbering the other groups: the breakdown of the respondents in the study for each of the seven classes are provided in the Breakdown column. The 67 responses were all acceptable: that is, no entry seemed careless or illogical (e.g., simultaneously checking the like and dislike boxes), and no missing responses were found throughout the questionnaire.

Evaluation of Performance-Based Assessment

Question (hereafter, 'Q.') 1 in the slider format asked about their overall course experience and generated a mean score of 83.67; that is, their perception of the value of this particular learning experience was highly positive and welcomed. Table 4 summarizes the results of Qs. 21–22 asked them to assess the validity of the evaluation methods for the audio and writing assignments: students also revealed a high level of support for these methods. The two students who chose the 'other' option left similar comments suggesting they wished to make the 'three-point-rating scale' a 'five-point-rating' one, which does not specifically indicate that they regarded the rating itself as 'negative'. In summary, the feedback was positive regarding the PA-centered course design, and the level of acceptance of the semi-analytical PA assessment was also high.

Change in 'Can-Do' Notions

Figure 3 presents the results of Qs. 13–18 in the slider format, which, using the slider bar format, asked students whether their proficiencies in each skill had improved. The zero at the leftmost corner of the slider indicates no improvement as per students, while 100 at the rightmost corner corresponded to a high level of improvement. Different students have different notions of improvement; nonetheless,

Table 4 Students' evaluation of the evaluative methods

	Frequency	Percent
Q. 21 Audio assignment evaluation		
Good	60	89.6
Bad	5	7.5
Other	2	3.0
Q. 22 Writing assignment evaluation		
Good	64	95.5
Bad	3	4.5
Other	0	0

Table 3 Respondents' profile of students in the study

CEFR	Target students	Majors	Total	Breakdown	Percentage (%)
A2	Freshmen	Mathematics/Physics/Chemistry	27	13, 2	40
B1	Sophomores	Mathematics/Chemistry	21	6, 6, 9	31
B2	Sophomores	Mathematics/Applied Chemistry	19	16, 3	29

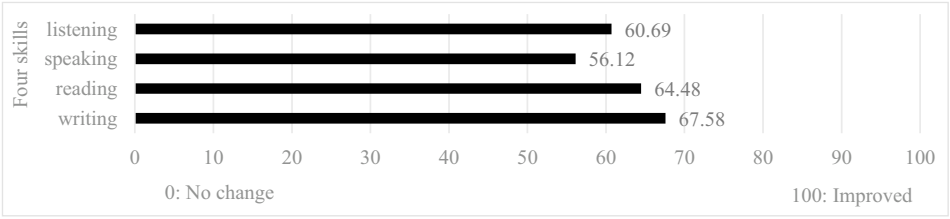


Fig. 3 Students' notions of improvement in the four skills

an average of 62.22 points toward the positive would indicate that students perceived an improvement in their overall skills, which cannot be ascertained in standard classrooms in Japanese culture. Among the four skills, the relatively low evaluation of speaking skill improvement may be attributed to the weakness in the simulative feature of the oral assignments in the course; conversely, students believed that their writing ability had primarily improved.

Figure 4 summarizes Q. 19 in the choice format about students' perceptions regarding their change in specific areas of English use after the course experience. The figure on the top collects all the items selected from multiple responses showing a positive perception, whereas the bottom notes the negative perceptions. Although several other factors need to be considered, the graphs reveal the students highly positive attitude toward different areas of English use; however, a small number of students suffered from a negative attitude.

Table 5 summarizes the TOEIC results coincidentally available (in Table 3, the Applied Chemistry major class), corresponding to approximately 30% of respondents. Among them, 16 students consented to use the data for co-analysis and triangulation of the survey data. The TOEIC tests were held twice toward the end of spring and fall semesters; incidentally, they formed pre-/post-tests—July and December—for the B2 class before and after the fall course of the present research. The total average scores increased by approximately 33 points from summer to winter. Paired-Sample *T*-Test confirmed that the reading scores—one of the targeted skills focused by the reading and writing textbooks used in the current study—have progressed ($p < 0.05$) though the sample number is small. Therefore, the improvement measured via the outer objective test medium also supports the students' perceptions about their improvement from the

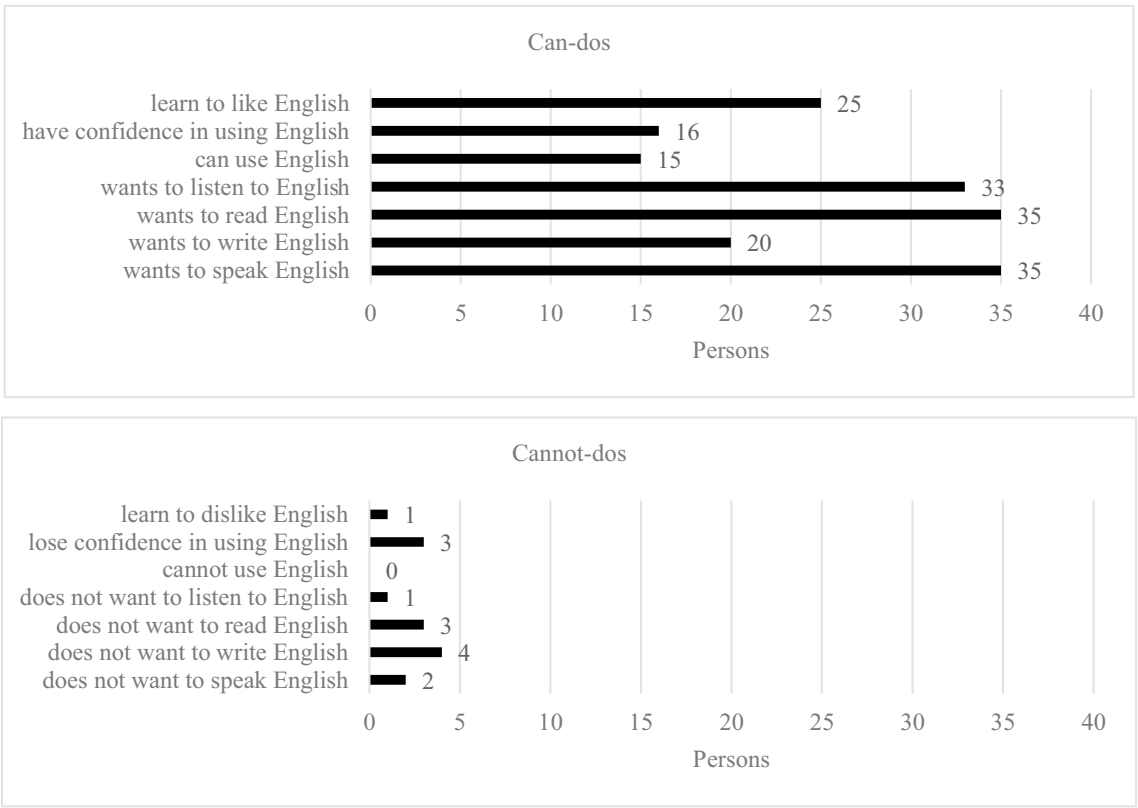


Fig. 4 Students' 'can-/cannot-do' notions after the course experience

Table 5 Change in the TOEIC scores of one B2 class

2018	*N*	Minimum	Maximum	Mean	Std. deviation
July					
Listening	16	140	320	248.75	42.249
Reading	16	180	280	233.13	31.563
Total	16	330	600	481.88	65.240
December					
Listening	16	170	330	257.50	43.589
Reading	16	190	330	257.50	39.073
Total	16	400	620	515.00	70.711

Table 6 Notions of improvement in the four skills by group

	Q. 13 Writing	Q. 14 Reading	Q. 15 Speaking	Q. 16 Listening
A2	64.63	65.30	65.22	64.78
B1	72.24	63.33	54.43	62.95
B2	66.63	64.58	45.05	52.37

survey: the students not only felt that they had improved, but they also had.

Relation Between Group Features and Assessments

A one-way ANOVA was executed to compare the means among the three groups (A2, B1, and B2) to determine whether they are significantly correlated to any specific question item. This process is needed to examine the possible effects of English proficiency of the respondents on their perception of the course evaluative design. Regarding Qs.13–16 (Table 6) on notions of improvement in the four skills (0: no progress, 100: improved), the test for homogeneity of variances as well as ANOVA significant values showed that the three groups were sufficiently homogenous for comparison, and the means were not considered to present a statistically significant difference ($p < 0.05$). In other words, the factors of initial English proficiencies in the study were unlikely to affect the outcomes on the PA instructional design; that is, the design would likely function at any level of English proficiency.

The same process was executed for Qs.1 and 17–18 (Fig. 5), which covered the students' overall evaluation of the learning experience using PA, as well as their preference ratios between exams and performance (0: test only, 100: performance only) if the students were to design their education, both in general and for the current course. All the group sets were considered statistically homogenous to enable mean comparison. Most importantly, the ANOVA showed that the mean scores among three groups for only Q.1 were statistically significant ($p < 0.05$); that is, B1—the middle-level group among three English proficiency categories—found the learning experience most meaningful, followed by A2 and then B2. Interestingly, all three groups considered the appropriateness of mixing approximately 15% of the test elements for the current instructional design. Contrarily, all groups considered that mixing from 20 to 25% of the test elements may be adequate for courses in general.

Finally, Fig. 6 presents a by-group analysis of Q. 20 in the choice format, which asked, 'If you became a teacher in the future, which evaluation method would you use?' Interestingly, an examination-only assessment policy was not considered the most effective by any three groups. It is noteworthy that no student in the B1 group wished to maintain the traditional 100% exam-only design that they would have been so much accustomed to; this tendency was more or less the same with the A1 and B2 groups.

Open-Ended Comments

Qs. 25–28 were open-ended. Q. 25 referred to the issue of subjectivity within educational measurements: 'Performance-based assessment is different from mid-term and final exams, but similar in that the teacher evaluates what students have produced. Please write any comments you may have on this point'. The question was phrased to avoid leading the students to favor one of the assessments as much as possible. Q. 26 asked students the most useful thing they learnt from the course, Q. 27 asked about any improvements that needed to be made in the future instructional design, and Q. 28 requested any further comments. Numerous comments

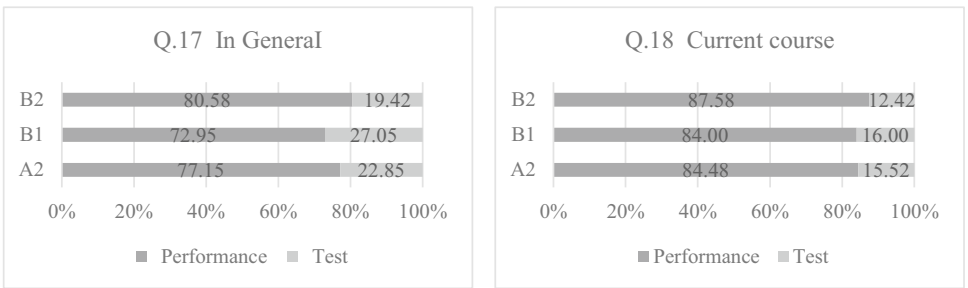


Fig. 5 Preference of instructional design by group

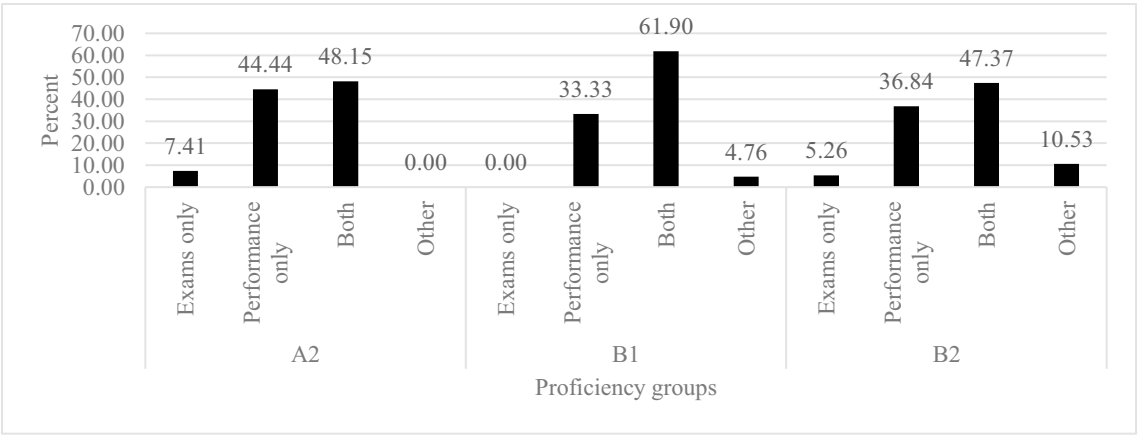


Fig. 6 Preference of assessment pattern as instructional design by group

were given, which made the application of AI text mining possible for Q. 25–46 (student comments), Q. 26–52, Q. 27–35, and Q. 28–39, excluding those who left a note such as 'nothing special'.

AI text mining produces the word cloud and hierarchical clustering representations for Q. 25. The four key concepts, 'performance', 'evaluation', 'strengthen', and 'accumulation', are highlighted, and the hierarchical clustering explains how these concepts are related in the comments. With a 10-line digest auto-generated by the AI, the PA seemed to positively reflect the day-to-day accumulation of efforts, although it could only vaguely clarify what point was scored. These findings seemed a fair and reasonable summative interpretation of the comments from Q. 25.

AI text mining also produces the word cloud and scored-word frequency of nouns and verbs in the comments for Q. 26. The highest frequency in the word cloud was the use of the terms 'can use', 'TOEIC', 'research paper', and 'speaking'. Scored-word frequency suggests the words of importance using the term frequency-inverse document frequency (tf–idf) statistical arrangement that characterizes the documents over the simple frequency of appearance [41]. From the 10-line digest, the students seemed to think that the course could be used to speak in English with foreigners, read and write research papers in English, and prepare for TOEIC tests.

In response to Q. 27, 12 out of 35 comments stated, 'It is fine like this', whereas the rest noted different points, particularly regarding class management. The AI included five lines out of 10 that said 'It is fine as it is' with slightly different phrasing, proportional to the 12 out of 35 comments. It also listed three specific points:

- An average class size of 20 is preferable.
- Some course exam evaluations should be added.
- Some additional assignments are fine.

These points for improvement were noted by single comments from three of the students; the AI seemed to consider these points significant and listed them in that manner. In the additional comments for Question 28, many students made polite remarks of thanks besides adding that the course was 'fun'.

Implication

This research examined the perceived effectiveness of PA in blended learning via student evaluations. The course was designed with a 100% PA policy, and its learning outcomes were assessed via a survey. Specifically, the research found (1) a high level of positivity toward the PA-centered evaluation methods (Table 4), (2) high notions of improvement ('Can-dos') covering all the four skills and specific areas in use (Fig. 3), (3) applicability of the design regardless of the students' initial English levels (Table 6), and (4) only a small number of students who wished the courses (both the current course and in general) to be 100% test-based (Fig. 6). The students' positive attitude toward the course design was further partly confirmed by the English proficiency test results (Table 5) drawn from outside the current research scheme for triangulation. Moreover, the literature review found that a 100% PA course design, as evaluated by the students, was non-existent.

Based on the literature review and these findings, five points are selectively discussed. The first point concerns students' self-confidence or 'can-do' notions. Students in Japan tend to underestimate their actual abilities, perhaps as part of a culturally estimated virtue [28]. Their uncertainty about their abilities cause them to interpret low scores as even lower: the can-do notion, that is, directing their attention to comparing their past status to the current to see the progress they have made, as was investigated via the survey questions,

may be beneficial for students as they begin to think, 'I can now handle it better than before', which is more accessible than if presented a digital number from pen-paper tests that would make them think, 'I am going to do worse on the next test'. Placing PA as a core to check their progress may be a better option in Japanese culture, particularly in subject areas such as language education, in which self-confidence is critical for production.

The second point relates to the potential use of PA-centered instructional design of blended type in the broader context of situations, such as the pandemic we faced. Education the world over is now facing an unprecedented challenge, namely, the near impossibility of sharing the same physical space for teaching and learning in a classroom. Consequently, numerous test events such as term examinations and some on-site English proficiency tests had had to be postponed or canceled. PA-centered instructional design can be a viable alternative in these circumstances as it allows teachers to evaluate without time constraints and the challenges of physical space sharing. For a long time, identifying students and their digital products has been a hurdle in online learning: the asynchronous voice recording type of homework as was applied in the current study could be a feasible solution, more effective if combined with synchronous online sessions in which spontaneous vocal response from attendees can be realized. If automated in an LMS, voice-based authentication of students' digital products [42] would be a simple but practical solution to numerous problems involved in online teaching and testing.

The third point is related to the ideal instructional design in terms of evaluation, as the Cizek study [22] had motivated. The results of the survey show that students in this study considered 15% of test elements for the current course design and from 20 to 25% of test elements for the course, in general, to be suitable. This result may signify that the students see the merits of objective tests after 100% of PA experience; furthermore, they may think the other courses weighing on tests could reduce the test elements extensively. Eventually, we may conceive an instructional design concept in which the combinatory ratio of PA and tests is variable, depending on the nature of course contents, curriculum structures, students' preferences and learning styles, and other factors. Students who regard test-only policy became near zero regardless of their learning stages (Fig. 6) after this study provides evidence of how important experiencing fully embedded PA course structure themselves, particularly for those who may become teachers in the future.

The fourth point regards the size of the online class community. The open-ended comments in the survey analyzed by the AI text mining produced a summative sentence: 'An average class size of 20 is preferable'. This comment echoes the author's prior doctoral thesis study, conducted in 2007–2008, which found that participants might feel the most robust sense of unity around a class size of 20 students in a blended classroom community, as Fig. 7 presents [44, p.85]. The study applied an assessment scale developed by Rovai to calculate the strength of the class community [44]. It is unclear why this specific student provided its feedback with a specific number of '20' as the ideal class size in blended learning. Refinement of parameters in instructional design, the issue of ideal—most effective and efficient in teaching and learning—would be another possible avenue to pursue in future trials.

The last point concerns our coexistence with an AI-powered society. AI played a vital role in database search and text mining in the current research. With the relevant keywords, the Scopus search produced closely related articles, including 'students' perspectives' in seconds, which would not be possible with a random search. On the other hand, the 3–5–10 line digests of the AI text mining produced a small number of odd statements or selected particular sentences that were unexpectedly highlighted. For instance, in the word cluster, the two expressions 'English' and 'English ability' used by the students could be interpreted as a sign that the students had begun to distinguish 'studying English' from 'acquiring the language'; however, the AI did not seem to differentiate between them. Nevertheless, all these processes would necessitate a lengthy period for collection and analysis without the aid of AI, requiring co-coders and co-researchers to verify and balance any human errors/bias that would naturally be involved. If used appropriately, AI can certainly be a helpful tool to offset these issues with larger data sets and limited researcher time.

This study admits limitations in its research approach. The respondents in this study were primarily science-major students: further data collection from students of different disciplines should ideally be made to counterbalance the

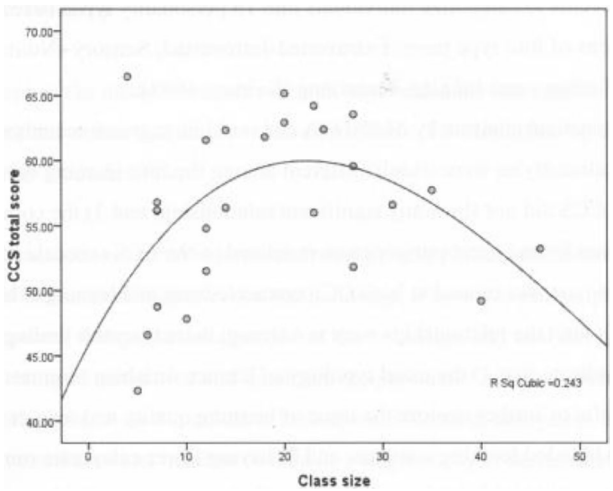


Fig. 7 Class size and Classroom Community Scale (CCS) total score

research findings. Also, the volunteer-based participation in the survey collected a relatively small sample size for higher statistical analyzes: further research with a much larger sample size with multiple teachers' cooperation who follow a similar performance-centered instructional design would be beneficial to gain further implications to assess the PAs. Finally, the experimental research design of test-only versus performance-only comparison with all the other variables to be the same as possible can be another avenue to be further pursued, if accidentally be designed, because this could risk research ethics of providing undesirable stimuli to one group of the students-subjects if the researcher-teacher believes performance-centered approach to be the most beneficial with less adverse effects to the students' learning.

Conclusion

This study demonstrates that the performance-based assessment course design is effective in helping students' learning process, regardless of their learning stage. Furthermore, the results suggest that the combinatory design of primarily performance-based with some test-based assessments will make the course more authentic, acceptable, and more relevant to students. The proposed PA-centred course design could potentially apply to several other fields where performance as outcomes is involved. This study appeals to its contribution in providing a concrete research example in reply to the research agenda proposed by Cizek [22]; namely, assessment should preferably be embedded in the instructional process to encourage and produce progress in students' learning. As the PA experience evidenced, online elements are an essential part of world education, especially once their merits are experienced. We need some more time to know what the ideal ratio of traditional face-to-face elements and online elements would be, but we believe we are ready to hear the students' voices if they wish it to be primarily online with some face-to-face elements, and not exclusively one of them, on the continuum of blended learning.

Acknowledgements The online parts of the course contents on the LMS were derived from the Cambridge University Press course textbooks with the publisher's generous permission. We also express our gratitude toward the Tokyo University of Science students, who provided their kind accord to the study's survey participation, analysis, and publication.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Doyon P. A review of higher education reform in modern Japan. High Educ. 2001;41(4):443–70.
2. Tomoda A, Mori K, Kimura M, Takahashi T, Kitamura T. One-year prevalence and incidence of depression among first-year university students in Japan: a preliminary study. Psychiatry Clin Neurosci. 2000;54(5):583–8.
3. Lee M, Larson R. The Korean 'examination hell': long hours of studying, distress, and depression. J Youth Adolesc. 2000;29(2):249–71.
4. Yu L, Suen HK. Historical and contemporary exam-driven education fever in China. KEDI J Edul Pol. 2005;2(1):17–33.
5. Cantu DA, Warren WJ. Teaching history in the digital classroom. Routledge; 2015.
6. Sweet D, Zimmermann J (eds.) Performance assessment. Education: consumer guide. 1993; n2 Nov. https://eric.ed.gov/?id=ED353329. Accessed 20 Dec 2021
7. Ministry of Education, Culture, Sports, Science and Technology, Japan (MEXT). Chapter 4: Direction of improving teacher skills in performance evaluation. 2017. https://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2017/10/04/1395572_02.pdf. Accessed 20 Dec 2021
8. Kawaijuku. Changing high schools: performance assessment. Kwaijuku Guidel. 2017;4–5(13):25–44.
9. Madaus GF, O'Dwyer LM. A short history of performance assessment: lessons learned. Phi Delta Kappan. 1999;80(9):688.
10. Elsevier. Journal title lists. 2021. https://www.elsevier.com/solutions/sciendedirect/journals-books/journal-title-lists. Accessed 20 Dec 2021
11. Glenn F. Language testing, assessment, and educational measurement journals. n.d. http://languagetesting.info/journals/list.html. Accessed 20 Dec 2021
12. Hambleton RK. The rise and fall of criterion-referenced measurement? Educ Meas. 1994. https://doi.org/10.1111/j.1745-3992.1994.tb00567.x.
13. Camara WJ, Brown DC. Educational and employment testing: changing concepts in measurement and policy. Educ Meas. 1995;14(1):5–11.
14. Cizek GJ. Standard-setting guidelines. Educ Meas. 1996;15(1):13–21.
15. Nichols P, Sugrue B. The lack of fidelity between cognitively complex constructs and conventional test development practice. Educ Meas. 1999;18(2):18–29.
16. Hambleton RK, Jaeger RM, Plake BS, Mills C. Setting performance standards on complex educational assessments. Appl Psychol Meas. 2000;24(4):355–66.
17. Koretz D. Using multiple measures to address perverse incentives and score inflation. Educ Meas. 2003;22(2):18–26.

18. American Educational Research Association (AERA). The standards for educational and psychological testing. AERA Publications; 1999.
19. American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). Standards for educational and psychological testing. 2014. https://www.apa.org/science/programs/testing/standards. Accessed 20 Dec 2021
20. Gotch CM, French BF. A systematic review of assessment literacy measures. Educ Meas. 2014;33(2):14–8.
21. Koh KH. Improving teachers' assessment literacy through professional development. Teach Educ. 2011;22(3):255–76.
22. Cizek GJ. An NCME instructional module on: setting passing scores. Educ Meas. 1996;15(2):20–31.
23. Klein SP, Jovanovic J, Stecher BM, McCaffrey D, Shavelson RJ, Haertel E, Solano-Flores G, Comfort K. Gender and racial/ethnic differences on performance assessments in science. Educ Eval Policy Anal. 1997;19(2):83–97.
24. Fox J, Cheng L. Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. Assess Educ Princ Pol Pract. 2007;14(1):9–26.
25. Barkaoui K. Rating scale impact on EFL essay marking: a mixed-method study. Assess Writ. 2007;12(2):86–107.
26. Kozaki Y. An alternative decision-making procedure for performance assessments: using the multifaceted rash model to generate cut estimates. Lang Assess Qual. 2010;7(1):75–95.
27. Kondo-Brown K. A FACETS analysis of rater bias in measuring Japanese second language writing performance. Lang Test. 2002;19(1):3–31.
28. Matsuno S. Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. Lang Test. 2009;26(1):75–100.
29. Penny JA, Johnson RL. The accuracy of performance task scores after resolution of rater disagreement: a Monte Carlo study. Assess Writ. 2011;16(4):221–36.
30. Kane MT. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50(1):1–73.
31. Wright KB. Researching internet-based populations: advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. J Comput Mediat. 2005;10(3):JCMC1034.
32. Norman G. Likert scales, levels of measurement and the 'laws' of statistics. Adv Health Sci Educ. 2010;15(5):625–32.
33. Jamieson S. Likert scales: how to (ab)use them. Med Educ. 2004;38(12):1217–8.
34. Miyazoe T. A study on OERs, MOOCs, and LMOOCs for university students: how do they perceive and use them. Educ Stud. 2018;2018(60):1–17.
35. Glaser BGS, Strauss AL. The discovery of grounded theory: strategies for qualitative research. Aldine Transaction; 1967.
36. Miyazoe T, Anderson T. Learning outcomes and students' perceptions of online writing: simultaneous implementation of a forum, blog, and wiki in an EFL blended learning setting. System. 2010;38(2):185–99.
37. Council of Europe. The CEFR levels. n.d. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions. Accessed 20 Dec 2021
38. Wiggins G. Educative assessment: designing assessments to inform and improve student performance. Jossey-Bass; 1998.
39. Christensson P. QR code definition. Tech Terms. 2015. https://techterms.com/definition/qr_code. Accessed 4 Apr 2022
40. Tokyo University of Science (TUS). Undergraduate students' profile. https://www.tus.ac.jp/info/foundation/gakubu.html. Accessed 25 Dec 2019
41. Leskovec J, Rajaraman A, Ullman JD. Data mining. In: Leskovec J, Rajaraman A, Ullman JD, editors. Mining of massive datasets. Cambridge University Press; 2014. p. 1–19.
42. Miyazoe T, Anderson T. Advancement in online education. In: Lin Q, editor. Voice interaction online. Nova Science Publishers, Inc; 2012. p. 39–67.
43. Miyazoe T. LMS-based EFL blended instructional design: empirical research on the sense of class community, learning styles, and online written interaction. Doctoral dissertation. International Christian University, Tokyo; 2009
44. Rovai AP. Development of an instrument to measure classroom community. Internet High Educ. 2002;5(3):197–211.