Nakata et al. Language Testing in Asia (2018) 8:15
https://doi.org/10.1186/s40468-018-0067-2

Language Testing in Asia

RESEARCH Open Access

Assessing Japanese teachers' classroom English "internationally": implications for the development of classroom English language benchmarks in Japan

Yoshiyuki Nakata[1*], Osamu Ikeno[2], Yuzo Kimura[3], Naoyuki Naganuma[4] and Stephen Andrews[5,6]

* Correspondence:
nakata.yoshiyuki@gmail.com
[1]Faculty of Global Communications, Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
Full list of author information is available at the end of the article

Abstract

Background: This study aims to develop a low-stakes assessment tool to establish a classroom English language benchmark that Japanese teachers of English can use for their own professional development purposes. To start with, we describe the differences between CLA (Classroom Language Assessment) in Hong Kong and the IDS (Integrative Diagnostic Scale) in Japan with regard to agenda, characteristics, and implementation. Then, we report findings made from both of these assessment types and from a group discussion we had in order to clarify the rationale behind the CLA experts' assessment.

Methods: The participants of the present study consisted of two groups: (1) four project members of the IDS in Japan and (2) two CLA raters in Hong Kong. Following an initial assessment of videotaped performances by the IDS developers (based on the IDS benchmark) and assessment of the same videotaped performance by the CLA experts (based on the CLA), a follow-up focus-group discussion between the IDS developers and the CLA experts was held.

Results: The results of the study offer some evidence to support the thesis that a classroom English benchmark helps capture aspects of classroom English language proficiency that are indispensable for students' language acquisition. Accordingly, on the basis of the findings, suggestions for further revisions of the IDS were made.

Conclusion: The study showed that, with some revisions, the IDS can be an apt professional development tool through which Japanese teachers of English can improve their classroom English proficiency. It also suggested that more research is needed to contextualize the IDS as such a professional tool more fully.

Keywords: Classroom English language proficiency, Benchmark development, Professional development, Secondary EFL teachers

Introduction

We are all aware that rich input or target language use promotes second language acquisition (Kim and Elder 2008; Macaro 2014; Tsui 1985). Teachers have a critical role in their students' language acquisition and are expected to be sufficiently fluent and proficient in the target language—here, English—to provide those who learn from them with ample opportunities to use it, interact with each other in it, and express

themselves through its grammar and vocabulary. It is for this reason that the Japanese Ministry of Education, Culture, Sports, Science and Technology (2010, 2015) is currently interested in English medium instruction of English language learning (i.e., EMI in an EFL context).

However, in East Asian countries like China, South Korea, and Japan, which are faced with rapid globalization, English medium instruction (EMI) has often been started not on teachers' initiatives, but in response to voices from industry and high-stakeholders, appearing under the guise of "teacher quality control." An example is the Classroom Language Assessment (CLA-HK hereafter) in the Language Proficiency Assessment for Teachers (English Language) (LPATE) in Hong Kong (Coniam et al. 2017). Japan, though lagging behind other countries in Asia (such as China and South Korea), is no exception to this trend (Freeman 2017). MEXT (2010) has issued course-of-study guidelines for high schools that includes the statement: "English classes should be taught in English" (p. 3). This trend is becoming even more prominent in Japan, which is hosting the Tokyo Olympic Games in 2020. Already through the implementation of this guideline, there has been increased use of English in English lessons both in high schools—47.9% in first-year grade (upper secondary school)—using the textbook "Communication English 1" (MEXT 2015).

The important question to be addressed here, therefore, is how we can help teachers seeking to improve the quality of EMI lessons in order to encourage their students to use the language they seek to learn and accelerate their language acquisition. Any benchmark with a specific focus on the assessment of teachers' classroom language could potentially serve as an apt means of making "a link between the test [i.e., the diagnostic outcome of teachers' classroom English] and the classroom objective [i.e., teachers' awareness raising about their EMI lesson and its efficacy on the improvement of EMI lessons]" (Chapelle and Voss 2014, p. 1092).

Hence, the goal of this present study is to improve the quality of the low-stakes (rather than high-stakes) assessment tool known as the Classroom English Language Proficiency Benchmark so that Japanese secondary teachers of English can utilize it for their own professional development purposes. More precisely, as part of the process of developing this tool, the study attempts to investigate how CLA experts in Hong Kong assess Japanese EFL teachers' English language proficiency in the classroom using the CLA-HK, and thereby to explore appropriate use of the benchmark Integrated Diagnostic Scale in Japan (IDS-J henceforth) and ways in which it might be further improved. The study is based on the assumption that the CLA-HK, from which the IDS-J was developed, can itself be an appropriate tool for examining the quality of the IDS-J scale. The involvement of CLA-HK experts in assessing teachers' performance helps us gain insights on which the IDS-J developers could build in order to enhance it.

Classroom English language proficiency and professional development

Perhaps, we all agree that the acquisition of such a professional skill as being able to conduct lessons in a foreign target language (EMI skills in an EFL context) does not happen overnight. Yet, the approach of how teachers acquire it is not without controversy. For administrators, it is a matter of teacher quality control (e.g., Teacher Quality Standards in Colorado, USA, and Alberta, Canada), often entailing high-stakes testing

which is mandatory for all non-native teachers (e.g., the LPATE in Hong Kong). Attempts to promote professional development in this manner could justifiably be seen as another unreasonable burden on teachers, who already have a heavy load of work and many duties over and above preparation for lessons and lesson improvement, if they are not given enough time for training and some degree of choice. For many (if not all) teachers and teacher educators, it is a matter of teacher-initiated professional development: the teachers themselves decide their professional development targets or skills, so that they can improve particular aspects of their general capability at their own pace (see Pasternak and Bailey 2004; Nakata 2011). This is in stark contrast to the externally molded kind of professional development requiring teachers to achieve a pre-ordained target skill level.

Professional development involving EMI inevitably entails a shift of teaching style from "a teacher-led" to "a more interactive dynamic one" and is thus challenging for teachers who have "limited self-taught or no previous knowledge of EMI in the class" (Dearden 2014, p. 6). Without EMI, however, teachers cannot create a secure classroom environment where students are willing to speak and interact with others in English. What matters most is *learner development* or *teacher development for the sake of learner development*, and this means that the real issues are how effectively in practice can teachers acquire the skills to provide their students with opportunities to use the target language, and how can they continually improve? If these issues are disregarded, both students as learners and teachers seeking to become better practitioners will fail to reap the maximum fruits of EMI practice, and thus, the positive effects of EMI will be limited.

For language teachers, the specialist language skills required are "command of subject specific/metalinguistic terminology" and "the discourse competence required for effective classroom delivery of subject content" (Elder 2001, p. 152). The assessment tool must be able to diagnose such language proficiency aspects (see Alderson et al. 2015 for diagnostic language assessment). The developmental process also needs to encompass such elements as "inquiry into the meaning of test [i.e., classroom English language proficiency assessment] scores, their use, and their consequences [i.e., teachers' awareness raising about their EMI lesson]" (Chapelle and Voss 2014, p. 1082). Moreover, such practices are likely to be different across teaching contexts (Chapelle and Voss 2014) and even more so in the case of low-stakes tests.

Taking all this into account, a benchmark essentially needs to be developed through a series of developmental processes of low-stakes (rather than high-stakes)[1] tests including "objective reflection in which the [course] objectives are reflected by the test tasks and the abilities that it measures" (Chapelle and Voss 2014, p. 1089).

Only through a process like that described—with a suitable benchmark—can a low-stakes diagnostic tool be developed which measures the specific language needs of the classroom situation. (See Douglas 2001 for English (Proficiency) for Specific Purposes (ESP or EPSP) assessment as distinct from general language proficiency assessment).

The CLA-HK as an assessment tool for teachers' English language proficiency in the classroom

As is often the case in East Asian countries, the Hong Kong secondary school EFL context is more or less similar to the Japanese one. Teachers have large classes of, say, 35–45

students and have to maintain a specific focus on grammar, vocabulary, and reading as a preparation for an entrance exam (Tsui 1996). Yet, in the eyes of the IDS developers, language education and language teacher education[2] in Hong Kong is much more advanced than in Japan (see Education Bureau 2017).

The Language Proficiency Assessment for Teachers (English language) in Hong Kong (from henceforth the LPATE) appears to offer important insights into the English proficiency required of classroom teachers (Coniam and Falvey 1999; Coniam et al. 2017; Sewell 2013). The LPATE is the language assessment test officially adopted for assessing the linguistic competence of EFL teachers in Hong Kong primary and secondary schools and is now implemented by the Hong Kong Examinations and Assessment Authority (the HKEAA) and the Education Bureau (EDB) there. The LPATE requires all EFL teachers to meet a minimum proficiency standard, either by passing its benchmark tests or by attending accredited training courses. The language proficiency of the teachers who participate is assessed by the LPATE through three pen-and-paper tests (reading, writing, listening), a speaking test, and Classroom Language Assessment (the CLA) (see EDB (2017) for more details).

Unlike the pen-and-paper tests, the CLA-HK is performance-based. The assessment of the candidates' English language proficiency,[3] as teachers in the classroom, is made by two assessors over a continuous period of more than 20 min' observed teaching, twice on different days (see the LPATE handbook (HEKAA 2011) for more details). This is based on the CLA-HK, which has a five-point rating scheme for each of the following aspects:

Scale 1: Grammar.
Scale 2: Pronunciation, stress, and intonation.
Scale 3: Language interaction.
Scale 4: Language of instruction.

What makes the CLA-HK distinctive from other oral language performance assessments, such as the ACTFL (American Council on the Teaching of Foreign Languages) Proficiency Guidelines (ACTFL 2012; see also Kimura et al. 2017), is the inclusion of scales 3 and 4, which guide assessors in how to "measure how appropriately a candidate can elicit students' answers, respond to them, provide them with feedback, give them instructions, and give them signals in his or her lesson when necessary, taking learning proficiency level into account" (Nakata 2010, p. 78). As such, the CLA-HK, when implemented appropriately, can be an apt or preferred means of capturing the multifaceted nature of teachers' classroom proficiency in promoting students' English use.

However, as is often the case with benchmark systems, the CLA-HK was criticized at an early stage of its development, for lack of significant contextualization due to overhasty piloting and insufficient time allowance (Dawson et al. 2003; see also the Canadian Language Benchmark[4]). The current version of the CLA-HK has been developed after experimentation lasting more than a decade and with subsequent revisions by the HKEAA and EDB.

In the knowledge that the CLA is a high-stakes' assessment, the Nakata (2010) of this article examined its potential as a professional development tool for EFL teachers in Japan and concluded that "the development of the Classroom Language Assessment

benchmark for English teachers should be based on the premise that English teachers should use it as a reflective, interactive, and pedagogical tool" (p. 89). It was found that Classroom English observation using the CLA-HK Benchmark and the CLA-HK Assessment Sheet is effective "for raising trainees' awareness of their classroom English and helping them strive for improved classroom language proficiency" (p. 76). This is the case where CLA-HK developed as a high-stakes test is used as a low-stakes test for the purpose of teachers' professional development. Moreover, this review also suggested that "an improved, simplified benchmark and assessment sheet should be developed so that when used together they would serve as an important means for language teachers' professional development" (p. 76). As Dawson et al. (2003) argue, what is most important in the developmental process is its contextualization (in this case, in the Japanese EFL educational context). Hence, such a benchmark must be "developed and revised for the better through a trial-and-error process that will enable EFL teachers to use it collaboratively as a professional development tool" (p. 89). Following this suggestion, an Integrated "Diagnostic" Scale (IDS-J) was developed for Japanese teachers (Nakata et al. 2012), based on the CLA in the Hong Kong LPATE, but for a different purpose, namely teacher-initiated professional development.

Developing classroom English language benchmarks in the classroom

What would be the key characteristics of a set of English language benchmarks for the classroom that could potentially contribute to teacher-initiated professional development in the EFL school context? They would need to satisfy the following six points:

1. The primary objective of the set is to promote the professional development of EFL teachers; it is not assessment for its own sake.
2. The set of benchmarks is used as tools to promote teachers' awareness regarding their classroom English.
3. The focus of assessment is on the aspects of EFL teachers' language proficiency that relate to effective language use in the classroom.
4. Proficiency in the target language is rated as being inextricably linked with, although separable from, instructional skills and subject matter knowledge.
5. The set of benchmarks is specifically adapted for EFL teachers (in this case, EFL teachers in Japan), many of whom adopt the teacher-centered approach of Presentation, Practice, Production (PPP) (Shintani 2016).
6. The benchmarks can be used in post-lesson observation conferences in schools and graduate programmes and in in-service teacher development seminars by the Board of Education.

The IDS-J was developed, together with the other three scales, in an attempt to help teachers know and be responsive to the benchmarks that meet these requirements. As it was developed from the CLA-HK, it can be assumed that the IDS-J and the CLA-HK share a basic underlying concept (Ikeno et. al. 2016; Kimura et. al. 2017; Nakata et. al. 2012).

Given this common foundation in basic principle, how is the IDS-J[4] different from the CLA-HK? There are several salient characteristics of the IDS that are distinct from those of the CLA. First, unlike a high-stakes assessment tool such as the CLA in the

Hong Kong LPATE, whose purpose is to control the quality of language teachers (see the handbook of LPATE for more details), it has been designed to assess/evaluate the overall level of teacher language proficiency in English lessons. For this reason, it is observation-based—it includes appraisal of videotaped performance—and assessment can mainly be carried out by external evaluators. The evaluators' comments are used as informational feedback to the teacher being assessed, perhaps with some discussion about his or her lesson in the post-observation conference. Contrastingly, in the CLA-HK, due to its assessment purpose, no connection between the candidate and the assessors is allowed other than in observation, and thus, there is no post-observation conference.

The second salient element of the IDS-J is that it presents a single scale, though there are five perspectives or assessment areas (Fig. 1):

Grammar (accuracy and variety).
Vocabulary (appropriateness and variety).
Pronunciation (accuracy and naturalness).
Instruction and Explanation (efficiency and clarity).
Interaction with Students (smoothness).

The IDS-J assumes that the target lesson includes all of these five elements, in a similar way to the CLA-HK assumption of there being four elements. As the IDS-J was modeled on the CLA-HK, there are some similarities between the two, especially as concerns *Grammar and Lexical Accuracy and Range* in the CLA-HK, which is further divided into *Grammar* and *Vocabulary*. After a number of discussions with research collaborators (mostly high school teachers), we came to agree that grammar and vocabulary should be assessed separately in the Japanese EFL school context, as this would be more likely to assist teacher development.

The third marked difference between the IDS-J and the CLA-HK is the type of target lesson the assessors aim to observe.[5] In the case of the CLA, the candidate is expected to conduct a student-centered lesson including four elements—(1) *Grammatical and Lexical Accuracy and Range*, (2) *Pronunciation*, *Stress*, *and Intonation*, (3) the *Language of Interaction*, and (4) the *Language of Instruction*. The lesson must show ample student-student interaction as well as teacher-student exchanges, with a wide variety of questions of both yes/no and display/referential types. For its part, the IDS-J allows teachers to conduct a normal lesson so long as it contains the five target elements just

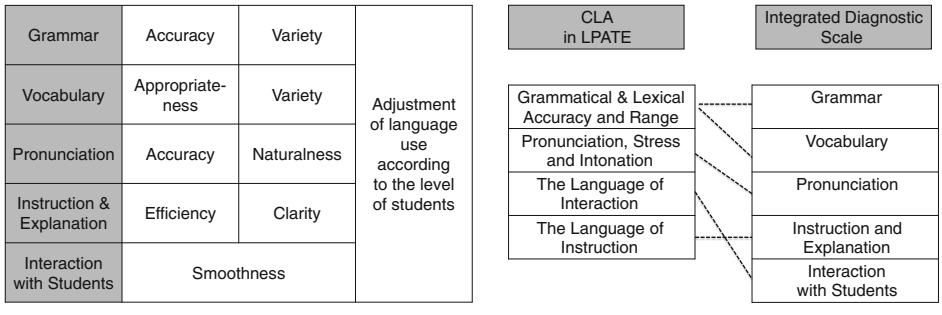

Fig. 1 Five scales (IDS)

mentioned. This is appropriate, since the purpose is for teachers to use the benchmarks for their own professional development in a given school context (often against considerable constraints and limitations).

The fourth feature of the IDS-J is that it has four levels of attainment, as distinct from the CLA-HK in the LPATE, which has five levels[4] (see Fig. 2). Employment of four levels was decided on after some field testing and discussions with secondary school teachers. The feedback from this was that having four levels was more practical and feasible for teachers pursuing their own professional development. A five-point grading seemed too detailed and made it harder for practitioners to discern where they were situated on the scale, especially in the area above and below the "pass" benchmark level (see *LPATE Handbook* (p. 81) for details). In the IDS-J, level 2 is the basic benchmark level, showing an acceptable level of language proficiency in the teacher (see Appendix and Additional file 1 for details). This is also a level from which assessments can go up to "good, proficient" or down to "not yet acceptable." Furthermore, level 4 in the IDS-J does not necessarily indicate fluency corresponding to level 5 in the CLA-HK, which implies a near native-like proficiency attainment.

The CLA-HK is more suitable as a high-stakes assessment tool, as it has more gradated intervals across its five levels and thus facilitates clearer distinction, even between the levels below the basic benchmark (levels 1 and 2). With this fit-for-purpose assessment tool, it should be possible for administrators to ensure the minimum quality of language proficiency required for language teachers in Hong Kong. However, the IDS-J is a more appropriate tool for assisting teacher-initiated professional development in its content and usage. For example, in the description of level 1: "Not yet acceptable," with its implication that all teachers are trying to improve their proficiency though, at this level, their effort has not yet paid off. Teachers are thus encouraged to do more to reach the basic benchmark (level 2 within their attainable level; also see Appendix), regarding it as a passing point for their life-long professional development. And once they have passed this level, they are likely to strive for even higher goals (levels 3 or 4) on their own initiative. In other words, this is a matter of their self-efficacy beliefs "I can do it" as well as helping with career development (Brophy 2010; Zimmerman 1998). This wording within the IDS-J scoring system is therefore certainly more encouraging—and even motivating—for those teachers who obtain a disappointing score but are well aware of their need for improvement.

The final salient feature, which should not be ignored, is how assessment is actually carried out in the IDS-J and the CLA-HK. With the CLA-HK, if there is "2" in any one

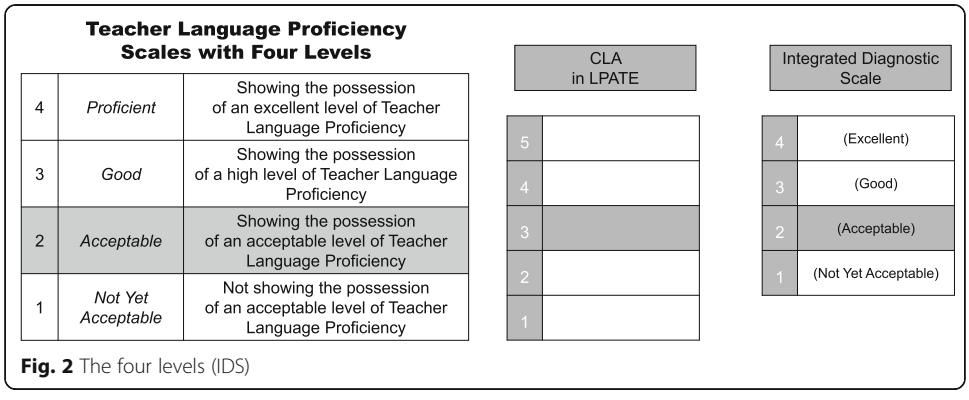

Fig. 2 The four levels (IDS)

area, the overall level is rated as "2," whatever scores the candidate has achieved in other areas (see candidate C in the CLA in Fig. 3). In the IDS-J scoring system, however, the attainment of an overall level of "X" requires "X" or an upper level ("X+1" or "X+2") in at least four of the five constructs and "X−1" is only allowed in one construct. For an example, see candidate B in the IDS in Fig. 3. Because the assessor gives "3" or "4" in four constructs but "2" in one construct, the overall level of candidate C is "3." Unlike the CLA-HK in which the lowest score reached in any area is the overall score, the IDS-J allows a candidate who obtains "X−1" in one area to be given the overall score of "X".

Now, let us take a brief look at each assessment category in the IDS-J. *Grammar* is assessed in terms of accuracy and variety as well by language use according to the level of the student. For example, level 2 in *Grammar* is: "uses grammar appropriately and the formation or self-correction of occasional errors does not impede the flow of the lesson as a whole." From this benchmark level, the score can go up to level 3 or 4, or go down to level 1. And in *Vocabulary*, the process is the same. Vocabulary is assessed in terms of appropriateness and variety as well as by language use according to the level of the student. In *Pronunciation*, accuracy and naturalness are the key features in the evaluation, but pronunciation must also be "internationally intelligible" (see Sewell 2013 for more details)—in other words, intelligible not only to "native speakers of English or those who share the same mother tongue" but also to "non-native speakers who do not share the mother tongue." Moreover, from the perspective of *English as an international language* (EIL), "EIL intelligibility" and "comprehend-ability" is closely related to "the ability to adapt to speaking and listening to EIL speakers of many stripes" (McKay and Brown 2016, p. 92). *Instruction and Explanation* are assessed in terms of efficiency and variety. Lastly, *Interaction with Students* is assessed with regard to "smoothness."

As we argued elsewhere (Kimura et. al. 2017), there remains another important procedure inherent in the development of a low-stakes assessment tool such as the IDS-J, and this needs further explanation. The additional feature is the field testing. It can be done both quantitatively and qualitatively, as explained below. The quantitative evaluation in the IDS-J can be carried out in the same way that experienced CLA-HK assessors evaluate videotaped performances in their system. The qualitative evaluation can be done through focus-group discussion between CLA-HK assessors and the IDS-J developers, with specific focus on the natures of the two systems, taking into account the difference between their two scales and the rationale for the gap between their assessment of quality, and allowing for adjustment of the IDS. Considering the social nature

CLA in LPATE

Rule

Candidates are judged as Level *X* when they obtain *X* (or a higher grade) at least in three of the four aspects and *X-0.5* only in one aspect, if ever.

	Candidate A	Candidate B	Candidate C
Grammar & Lexical Accuracy and Range	4	3	3
Pronunciation, Stress and Intonation	4	2.5	2
The Language of Interaction	3	3	3
The Language of Instruction	4	3	3
Overall Level	**4**	**3**	**2**

Integrated Diagnostic Scale

Rule

Candidates are judged as Level *X* when they obtain *X* (or a higher grade) at least in four of the five aspects and *X-1* only in one aspect, if ever.

	Candidate A	Candidate B	Candidate C
Grammar	4	4	4
Vocabulary	4	4	4
Pronunciation	3	2	3
Instruction & Explanation	4	3	2
Interaction	4	4	2
Overall Level	**4**	**3**	**2**

Fig. 3 The overall levels (IDS)

inherent in language assessment (McNamara 2001, p. 333), the stability of the benchmark can be enhanced only if it goes through the whole process. Once it is fully established, the system could perhaps be used more widely by teachers as a professional development tool.

Methods

Research purpose

The ultimate goal of this present study is to gain insights into the improvement of IDS-J in terms of revising the descriptors in the scale and identifying points which require consideration in its use. More specifically, the study attempts to achieve three aims: (1) to investigate how CLA-HK experts assess Japanese EFL teachers' English language proficiency in the classroom using the CLA; (2) to examine the reason for any rating differences of the same performance between CLA-HK experts and IDS-J developers, whose benchmarks are conceptually similar in essence; and (3) to explore appropriate use of the benchmark IDS-J and how it might be improved.

In order to achieve these basic aims, we followed a set procedure. As IDS-J developers, we made an initial assessment of each teacher's videotaped performance, based on the IDS-J benchmark. Then, the CLA-HK experts assessed the same videotaped performances according to their own CLA-HK scale. Following this, we held a focus-group discussion involving both the IDS-J developers and the CLA-HK experts. In this, we aimed to clarify the rationale behind the CLA-HK assessors' ratings, the process by which they were reached, and the reason why there might be any differences from the ratings given by the IDS-J developers (if such differences arose). By assessing the same videotaped teacher performances using the two scales—differing, though they shared the same basic concept—we also aimed to achieve some degree of qualitative (rather than quantitative) evaluation of our own benchmark IDS-J. We wanted to see whether Japanese teachers' classroom English was equally evaluated by the Japanese assessors and by the non-native English assessors whose mother tongue was a different one. We regarded this part of the discussion as essential to our enterprise, not only as authors, but also as IDS-J developers, as it would help us become more reflective about further possible revisions of the IDS-J.

Participants and context

The participants of the present study consisted of two groups: (A) the project members (raters A to D) and (B) two CLA-HK assessors (raters E and F). Details are given below:

(A) The team of project members[5] consisted of university professors with more than 10 years' experience as researchers and with some general knowledge about the development of benchmarks as well as about the CLA and the LPATE in Hong Kong (see Table 1).

(B) The CLA assessors were two university professors from Hong Kong with more than 10 years' teaching experience. While teaching at their universities, they have also served as assessors for LPATE (including CLA) for 6 or 9 years. As LPATE examiners, they had rated more than 1000 cases, and as CLA-HK assessors, they had rated between 200 and 300. They were qualified LPATE assessors authorized

Table 1 Details of the participants in the focus-group discussion

	Participants	Experience	Research interest
IDS developers	A	14 years' experience as a teacher educator (mostly for in-service teachers of English)	Motivation, self-regulation, teacher-learner autonomy
	B	18 years' experience as a teacher educator (mostly for pre-service teachers of English	L2 reading
	C	18 years' experience as a teacher educator and as a teacher researcher and 5 years' experience as a former high school teacher	Teacher motivation, dynamic systems theory
	D	14 years' experience as an EFL teacher	Can-do statements, CEFR
CLA assessors	E	9 years' experience as LPATE assessor, 10 years' teaching experience at the Teacher Training Institute in Hong Kong, 6 years' experience as an EFL teacher at secondary school	Language assessment, grammar teaching/learning, language across the curriculum
	F	6 years' experience as LPATE assessor, 13 years' teaching experience at the Teacher Training Institute in Hong Kong, 4 years' ESL teaching experience in primary school	Pronunciation, assessment, curriculum design

*The authors consist of the IDS developers above and an expert (the fifth author) in the field of teacher education and LPATE in Hong Kong

by the HKEAA and EDB and agreed to participate in the present research, making themselves available for the focus-group discussion. Their assessments, recorded below, are in many instances consistent with the representative ones described in the Assessment Report (see Education Bureau 2017, pp. 11–14).

So that the focus-group discussion can be contextualized, details of the participants are summarized in Table 1.

Material

Eight videotaped EMI performances were chosen by the project members based on the following criteria: (1) they had been widely accessible to many Japanese EFL secondary school teachers and had been used for teacher training sessions; (2) they could be used to assist the professional development of teachers; and (3) they brought up important points for discussion and for comparison with other EMI performances. The eight selected videos were edited down to 5 min prior to the assessment, in light of the elements included in the CLA-HK and the IDS-J.

Data collection

First step: assessment in Japan

Four IDS developers were asked to rate the eight videotaped EMI performances based on the IDS-J. They had several opportunities to watch the DVDs before the meeting and their final judgements were made in the meeting, where they watched them once again, discussing and reaching agreement on overall scores for each performance (see Table 2). This assessment was carried out in a quiet room with audio equipment that was available in a Japanese university. Having gone through the evaluation process based on IDS-J, the IDS developers were ready to interpret the findings in the next step, with the help of an expert in teacher education and the LPATE (Andrews 2007).

Table 2 Summary of the ratings

Performers	IDS developers						CLA assessors			
	IDS 4-point scale format						CLA 5-point scale format			
	Scales	AOS	A	B	C	D	Scales		E	F
Case 1	G	3	3	3	3	3+		G & L	3	3
English	V	3	3	3	3	3+				
Communication I	P	3	3	3	3	4		P	**2**	**2**
1st year	I & E	3	4	3	2	3+		Ins	3	3
AOS 3 (75%)	Int	2+	3	2	2	2+	**2 (40%)**	Int	3	3
Case 2	G	4	3+	3+	4	4		G & L	3	?
English	V	4	3+	3+	4	4				
Communication II	P	3+	3+	3	4	4		P	2	2
2nd year	I & E	4	3+	3+	4	4		Ins	3	?
AOS 4 (100%)	Int	4	4	3	4	4	**2 (40%)**	Int	3	?
Case 3	G	3+	3+	4	4	4		G & L	4	5
English	V	4	4	4	4	4				
Communication II	P	4	4	4	4	4		P	4	4
1st year	I & E	4	4	4	3	4		Ins	4	4
AOS 3+ (88%)	Int	3	3	3+	3	3	4 (80%)	Int	4	5
Case 4	G	2+	2	3	3	2		G & L	3	3
English	V	2+	2	3	3	2				
Communication Basic	P	2	**1+**	3	2+	2		P	3	3
1st year	I & E	3	**1+**	3	4	2		Ins	4	3
AOS 2+ (63%)	Int	N/A	2	2	3	2+	3 (60%)	Int	3	3
Case 5	G	2+	2+	3	2+	2+		G & L	**2**	**2**
English	V	2+	2	3	3	2+				
Communication I	P	1+	2	2	2	2+		P	**2**	**2**
1st year	I & E	2+	2+	3	2+	3		Ins	3	3
AOS 2+ (63%)	Int	2	2	2	2	2+	**2 (40%)**	Int	3	3
Case 6	G	3+	3+	3+	3+	3+		G & L	3	3
Reading	V	3+	3+	3	3+	3+				
	P	3+	3+	4	3+	3+		P	3	3
3rd year	I & E	3+	3+	3	3+	3+		Ins	3	3
AOS 3+ (88%)	Int	3	3	3	3	3	3 (60%)	Int	3	3
Case 7	G	3	3	3	3+	3		G & L	4	3
English	V	3	3	3	4	3				
Communication I	P	3	3	3+	3+	3		P	3	3
1st year	I & E	3	2+	3	4	3		Ins	4	3
AOS 3 (75%)	Int	2+	2+	2+	3	2+	3+ (70%)	Int	4	4
Case 8	G	2+	2+	2+	3	2+		G & L	3	3
Reading	V	3	2+	3	3	3				
	P	2+	2+	2	2+	2		P	3	3
3rd year	I & E	3	3	3	3	3		Ins	3	4
AOS 2+ (63%)	Int	2	2	2+	2+	2	3 (60%)	Int	3	4

Italics: subject name/bold: below the benchmark level
AOS agreed overall score/IDS: *G* Grammar, *V* Vocabulary, *P* Pronunciation, *I & E* Instruction and Explanation, *Int* interaction with students/CLA: *G & L* Grammar and Lexical Accuracy and Range, *P* Pronunciation, Stress, and Intonation, *Ins* the language of Instruction, *Int* the Language of Interaction

Second step: assessment and focus-group discussion in Hong Kong

Two experienced CLA-HK assessors were asked to rate the same eight EMI performances that the IDS-J developers had assessed earlier, but now basing their ratings on the CLA-HK. In order to delve deeper into the issues underlying the ratings and the value of the IDS-J, both the two CLA-HK assessors and the four IDS-J developers met for a focus-group discussion. For the assessment and the focus-group discussion in Hong Kong, the CLA-HK assessors granted permission for a recording to be made of what was said, as the IDS-J developers had requested. At the beginning of the session, raters E and F were asked about the CLA, the experience of the CLA assessors, and how the CLA-HK was being used in teacher education. Following this, the assessments and the focus-group discussions were conducted for each videotaped performance, one by one, in the following manner:

1. They viewed a given videotaped EMI performance.
2. They made their individual assessments of each performance based on the CLA-HK descriptors.
3. They discussed the overall scores they had settled on.
4. They gave an explanation of their assessment in a subsequent Q&A session.

The assessments of the eight EMI performances made by the IDS developers and the CLA-HK assessors are summarized in Table 2.

The final part of the session was a full explanation of IDS-J given by its developers followed by questions and answers. All this covered the aim of developing this assessment method, the process of development, the content, and perceived differences between the IDS-J and Hong Kong approaches.

Third step: content confirmation

Once the data in the discussion outlined above was collected and analyzed, a draft report was sent to the two experienced CLA-HK assessors (raters E and F), asking them to ensure that the report accurately reflected their intentions and that the information about the CLA and the LPATE was correct.

Fourth step: benchmark revision

On the basis of the findings, possible revisions to the IDS-J were discussed by the IDS developers, and suggestions for its further improvement both in content and usage were presented (see Appendix).

Results

Assessment in Japan

On the whole, two interesting features can be observed. First, the IDS-J developers gave higher evaluations in some fields, particularly on the performances of the teachers in cases 2 and 3. It is striking that raters C and D gave almost full scores to case 3's performance. The second salient feature is that, clearly distinct from other members, two project members (raters A and D) gave a relatively low judgment in case 4. Needless to

say, the rationale for these disparities needed to be moderated by subsequent CLA-HK assessors' ratings and scrutinized in the subsequent focus-group discussion.

Assessment in Hong Kong

Pronunciation, stress, and intonation: international intelligibility

The most significant feature in Hong Kong was that the Hong Kong assessors considered cases 1, 2, and 5's performances (i.e., agreed overall score) as "unbenchmarkable" or "below the benchmark level," as clearly distinct from the IDS developers who did not give an "unbenchmarkable" overall score to any of the performances they had observed. This means that three out of the eight EMI performers were labeled as "failures" or "unacceptable" at least by the CLA ratings.

This is not a surprising result, however, as the CLA-HK decisions were based on a rigid scoring system by which, if a candidate had a score below the benchmark level in more than one area, the total evaluation had to fall below the benchmark level. The common weakness across all the cases that fell in this way was *Pronunciation*, *Stress*, and *Intonation*, the raters maintaining that, if pronunciation was unintelligible, assessment could not proceed any further. For CLA raters who do not share the mother tongue with the videotaped teacher being assessed, pronunciation must be "internationally intelligible" and a fair judgment be given based on the CLA-HK descriptions of the teachers' proficiency levels. The following extracts are particularly vivid illustrations of why low scores were allotted in cases 1 and 2:

> We both agree that it's the weaknesses in pronunciation, stress, intonation which actually influence her other domains as well. Because we couldn't actually hear her clearly. We couldn't understand her fully. So we would not be able to actually suggest high marks for grammar and lexis, for example, or even interaction and for instruction as well. Definitely we would suggest a '2' for pronunciation, stress, and intonation. (Rater E, case 2, S4 122:54 recording time)

This kind of problem also came up with some weaker candidates who sat for the LPATE, as can be seen in the Assessment Report's CLA section:

> Weaker candidates did not place sufficient focus on clarity of individual sounds.
>
> Wrong words were stressed or every word in a sentence was pronounced with the same degree of stress, resulting in monotone.
>
> Pauses were not used effectively, leading to a breakdown in understanding.
>
> (Education Bureau 2017, p. 13).

Another defining feature was the divergence in the rating of case 2 between CLA and IDS. Both raters E and F gave a score of 2 (out of 5) for case 2's overall performance and for her pronunciation, which is below the benchmark level. This low score stands in stark contrast to the IDS developers' high ratings (mostly 4 and 3+). This is certainly an important point that needs to be explored.

As mentioned earlier, case 2's weakness lay in her pronunciation, which impeded intelligibility, and it was this that accounted for the CLA-HK score, it being an overall score as well. As one of the IDS-J developers, rater A, commented: "[Generally] it was an acceptable level … in our judgement … this, maybe, because we share the same language?" Rater F remarked: "It [sounds] very Japanese to me, just like my bunch of friends in [Japan]. Very gentle voice. And intelligibility is an issue here. That's why I can't mark the other [aspects]" (Rater F 124–125; case 2). The same rater also said: "I don't necessarily think it's the Japanese accent that causes the problem. Say, for example, … Case 7, she is actually speaking with a Japanese accent, but she can still get reasonably high scores for pronunciation, for grammar" (Rater E 131:220; case 2).

So, in contrast to the teachers in cases 4 and 7, whose pronunciation had some degree of Japanese accent but who were nevertheless intelligible (and thus benchmarkable), the teachers in cases 1 and 2 were found incomprehensible because of their pronunciation and could not pass the benchmark. The following remark from rater E clearly brings out the crucial point in judging a candidate's pronunciation:

> We don't penalize accents. We shouldn't. As long as she [with her Indian English accent] [remains] comprehensible, then that will still be okay. But if the accent influences people's comprehension of meaning, then that could be a problem. (Rater E 58:57)

Interaction with students: praise, feedback, and elicitation

Another problem area was *interaction with students*. For example, a point of criticism in case 1 was the teacher's excessive use of praise (one aspect of *Interaction with Students*), as can be seen below:

> Rater F: He used a lot of praise and encouragement, but no concrete feedback. Just like what Rater E said, this is a point that worries me a bit. … students cannot learn from anything. You say, "Excellent"; why "Excellent"? In terms of English language teaching, it is nothing. And they can't improve further. (142:37; case 1)

> IDS developer C: It's very interesting that you said that praise doesn't really work. Japanese English teachers often-- well, we are told generally that praising is a good technique for the students to develop. (144:24)

> Rater E: Yeah, it's not praising that causes the problem. It's actually just praising without concrete evidence. (144:50)

Though the case 1 teacher certainly tried to create a pleasant classroom atmosphere, his feedback was not necessarily helpful in raising students' language awareness. He did not make clear *why* they were praised and what was good. Rater E elaborated on this further, saying:

> [In] my experience of actually observing classes, successful interaction … is not just a kind of linguistic consideration; it's a kind of psychological consideration. The students know that they are not able to speak English. Just give them encouragement. Just try to show them you understand them instead of just saying,

> "Good, excellent." You say, "So do you mean...?" Sometimes actually showing this type of acknowledgement of the students being able to express meaning will be even more powerful … (Rater E 193:12) (my underlining)

The foregoing remarks from raters E and F were a response to the case 1 teacher, who lavished praise but largely failed to elicit utterances from his students. This failing can often be observed among weaker candidates of LPATE:

> inadequate ability in eliciting response from students or [failure] to react spontaneously to students' answers when required.
>
> Questions were repetitive or confined to those requiring one-word answers. While questions asked were mainly display questions, few attempts were made to ask extended questions, or to give hints or prompts when communication breakdowns occurred.
>
> (Education Bureau 2017, p. 14)

Here, we can see what potential there would be in utilizing this benchmark for a professional development purpose—to give feedback to the learners so they can improve their classroom English.

In the case above, the teacher, though trying to support the students' emotional confidence, failed to address their linguistic learning sufficiently; he missed opportunities to elicit student output and as a result failed to give his students full opportunities to use language in the classroom. Rater E (not only an experienced teacher trainer but also an experienced CLA-HK assessor) stressed the importance of "scaffolding" the students' opportunities to speak:

> We can't actually expect the students to be very articulate, especially in the English language, the first time they've got a question. But if they say even a word! Sometimes even in Hong Kong we've got very, very weak Hong Kong students speaking English. So they can [only utter] individual words, not connected together. Then the teacher could try to link up the words as uttered by the students … then follow this up with another question to try to elicit more and more responses from the students. (Rater E 193:12; case 1)

This may be a matter of teacher language awareness (Andrews 2007). More precisely, at the start, teachers need to be able to monitor their students' psychological and linguistic readiness as well as their needs, and adjust their language accordingly. In this way, they can come to help learners use the target language, eliciting or pinpointing individual concrete issues as they come up:

> Sometimes … we discussed the possibility of teachers overestimating or underestimating the students' needs, *et cetera*. It's a matter of how we modify the language as the lesson goes on. If the teacher, for example, does not receive responses from the questions, is he or she able to rephrase using simple language, or try to ask the same question in a different way? So we look at this kind of adaptation during class. It can be improvisation … at the top level of classroom language, they

> should be able to do this kind of improvisation according to the needs that they have just realized on the part of the students. (Rater E 198:14; no specific target) (my underlining)

This can be done only when teachers are aware of their own readiness to adapt approaches in their EMI classes. Thus, not only students but also teachers need encouragement to self-regulate their language use. The process involves analyzing students' needs (planning), teaching and giving feedback (performing), and evaluating the efficacy of the performance (reflection) (Zimmerman 1998).

Asked about the relation between the level of language the case 2 teacher used and their evaluations, the assessors denied a relationship, saying:

> I guess she's trying to prompt the boy to speak up. … So I'm not sure whether the question word here is right. Feedback is not enough. … partially because of the limited vocabulary in terms of feedback to the students, I think she needs [to do] some work on expanding the vocabulary, especially on the feedback to students. (Rater F 127:04; case 2)

Needless to say, both suggestions are insightful and could help the teacher progress in her professional development.

In either case, the common problem both the raters recognized in the teachers' performance was a lack of language proficiency with regard to *the language of interaction*, as was also pointed out in the Assessment Report as a recurring weakness (Education Bureau 2017): "The use of a restricted range of functional language was a common problem among weaker candidates." (p. 14).

Having heard rater 4's comment, rater E said, "I think we were biased by the difficult content she was dealing with in the textbook." This rater continued:

> I'm not sure if every … student could actually get what [some of the vocabulary] means, even if they could actually hear her. So in terms of the density of these specific terms, I would expect the teachers to try to elaborate a little bit more, maybe to cater for the needs of the weaker students in the class. So that even students who may not be able to get the terms the first time … could still benefit from the illustration elaboration. But I don't see this a lot in the video. (Rater E 133:25; case 2)

Case 3 is the only one in which both raters E and F gave more than 4 by the CLA descriptors. The following extracts give a fairly detailed account of the rationale behind their high rating in this instance:

> Rater E: Cool. So this one is much better, and I'm suggesting a straight 4. And she [Rater F] is suggesting a 5, 4, 5, 4 for this one. 5 would be the top level. Maybe you would like to discuss this?

> Rater F: Yeah, maybe I'll start first … Very natural use of language. And I like the way she used a lot of varied follow-up questions to prompt for more speaking opportunities. … She asked different people to respond to her. That created a lot of speech opportunities in class. … Language of interaction: '5' – because I can see it's a

> natural, authentic use of language. It's not a made-up one, but is so natural and genuine. … I'm afraid some people may not be able to follow because it's still a bit too natural, too spontaneous. So if we have some weaker students, maybe they're just shocked … not all of the students can follow what she's [saying]. (79:40-81:33; case 3) (my underlining)

> Rater E: Well, I would actually take this kind of lecture, this delivery, that's definitely a good point. I agree that sometimes not all the students could actually follow what the teacher is saying. But at the same time, if the teacher [slowed] down [her] speech so that everybody could actually listen to her, is that then the kind of model [of how] we would like our students to speak English? ... In terms of interaction, I did agree also that she was trying to interact with the students and trying to scaffold for the students, making the students to speak more and more. But still, the kind of responses that she is eliciting from the students would be rather brief. (81:34; case 3) (my underlining)

Striking here is the divergence of views between raters E and F with regard to the need for language modification in class—whether a teacher should modify language so that all the students (including the weaker ones) can understand or whether she should retain her natural English flow as a model. The final decision, whatever it may be, must be the one adhered to, so long as it is made through "exhaustive discussions" between two "experienced assessors" who are also teacher trainers. Through such meaningful discussion, raters can become even better professionals, both as raters and as teacher trainers and the quality of their assessment can be guaranteed. It is for this reason that assessors must be qualified and capable, not only as assessors but also as teacher trainers. To be equipped to give the right detailed suggestions, they must go through a number of assessor training sessions. Thus, the CLA-HK appears to be particularly successful when it comes to quality assessor training.

Teacher question style

An insightful observation from rater E can be seen in the underlined passage in the extract below.

> She was trying to interact with the students and trying to scaffold for the students, making the students … speak more and more. But still, the kind of responses that she [elicits] from the students could be rather brief. Say, for example, there was an occasion [to talk about] pizza: "I would like this". "I like pizza." And then the teacher says, "I like pizza too." Then she could actually build on this starting point and ask, "Why [do] you like pizza? What kind of flavour?" "I like tomato flavour" … she could actually develop this kind of potential to interact with the students, especially with the good relationship she [already has] with the students. (Rater E 81:34; case 3) (my underlining)

It brings out the need to shift teacher question style from the "I-R-F (Initiation-Response-Feedback) pattern" (Sinclair and Coulthard 1975) to a more "dynamic, dialogic pattern" in which teacher and students, then students and students, make those meaningful interactions that are more likely to accelerate second language acquisition. This

is based on the premise that students in the classroom do not necessarily acquire second language fluency from absorbing the teacher's "output" through EMI alone, but are more likely to do so through communication with significant others including their classmates (Bruner 1983; Tomasello 2003).

Here, we can glimpse what Japanese teachers of English have the potential to attain. With appropriate kinds of feedback about their classroom English from their assessors, they could become users of classroom English just as capable as the stronger candidates of LPATE (CLA), whose characteristics appear in the following comment on a teacher in action:

> Scaffolding using questions and cues worked well in encouraging various levels of response from students. Prompts were properly used to help students rectify their own mistakes and individualized feedback was appropriately given.
>
> (Education Bureau 2017, p. 14).

Systematic error and slips

There is a clear difference between systematic error and a mere slip (which even native speakers can make on occasion). The former is subject to rating, while the latter is not likely to be. Later in the group discussion, raters E and F considered the difference between systematic error and slips, pointing out that level 4 in the IDS-J does not make much of this issue, as does level 5 in the CLA-HK.

> IDS developer 3: I feel that the Hong Kong case is quite different … You're saying that 90% of your student English teachers can pass the CLA[-HK]. In our case [the Japanese case], the situation is much similar to, for example, South Korea or mainland China. (173:09)
>
> Rater E: Yeah, I agree. That's why I was saying that even if you're talking about Level 1 [in IDS-J] to cover some of the Level 2s [in CLA-HK] that we have just got, I wouldn't feel too surprised to learn that. But other than that, say, for example, Level 4 in [the] Japanese [IDS-J] scales would be … very close to Level 5 in [the] Hong Kong scale. (173:54)
>
> Rater F: [Y]our scale for Level 4 [in IDS-J] looks pretty much perfect [Level 5 in CLA-HK] to me. It seems like perfect, no errors. … Is systematic error allowed, acceptable? Intonation, sentence stress, word stress – it seems that they are not on the list. … In terms of pronunciation, you mentioned, "Almost always pronounces accurately and naturally." And speed, "Pauses." But then overall, … they [the candidates] are *generally* weak in intonation, sentence stress, and word stress. So will [these aspects] be part of [the assessment]? Or if they make minor mistakes, just like Case 3 – I [mean] just very, very minor mistakes – it should be a "4" in your case, definitely. (174:29)
>
> Rater E: In the Hong Kong case we say "with no systematic errors". There can still be slips of [the tongue, though], which is perfectly acceptable. (175:53)

This is further supported by a comment in the Assessment Report itself (Education Bureau 2017):

> "The ability to recognize and correct [students'] simple errors was evident among many of them [the LPATE candidates]. Although some grammatical errors were noted, communication was unimpeded on the whole" (p. 13).

This suggests that the IDS-J, as it is, has a wider range not only between levels 1 and 2 but also between levels 4 and 5 (see Kimura et al. 2017), as compared with the range in the CLA-HK (HEKAA 2011). Here, we can recognize possible room for further revision of the IDS-J—the description of level 4, in particular.

What makes level 5 in the CLA-HK?

In a similar and related vein, when asked about the characteristics of level 5 in CLA-HK, rater E gave a fairly detailed account of the qualities of performance expected, distinguishing them from level 4:

> The level 5 candidates just stand out … the students are interacting well with her [Case 3], and she's trying to elicit, making use of every single potential within the lesson to try to start interaction with the students, trying to be social, [showing] sensitivity to students' needs, adapting language, showing a range, sufficiently challenging the students … the students are actually engaged with language, not just with her as a teacher, but also with interacting with the teacher, using the English language. …to Rater F there are some qualities of '5' here. [But] … this is a little bit lower than '5' … So, if we are in doubt between a '4' and '5', we just say '4' [agreed overall score]… if we assess teachers for the CLA, we just look at the amount of interaction throughout the lessons as well as different stages of the lesson. We do expect teachers to interact with the students at all stages of the lesson, and try to grasp opportunities. (Rater E 83–84; case 3) (my underlining)

Type of lesson as an assessment target

As stated earlier, in the CLA-HK in the LPATE, the candidates need to prepare a student-centered lesson. This is antithetical to the concept of the IDS-J, which allows teachers more freedom as to the type of lesson observed. However, it may also be that when teachers stick to a teacher-centered lesson, they inevitably pay more attention to their own language and pronunciation than to their students' language use. In other words, the requirement of a student-centered EMI lesson in the CLA-HK may encourage teachers to speculate more about how to promote student language use than is the case when the IDS-J is used as a professional development tool.

> We usually stress that they have to plan a student-centered lesson. So … that the students are actually the core part. You have to plan what language they are exposed to, what language they are producing and are expected to produce. And your job as a teacher is to try to make your language comprehensible to them and try to facilitate their producing language … But the *way* that they speak: they are trying to engage the class – talking, interacting with the teacher or with each other – rather

than just keeping on explaining some difficult things to students. That would be clear evidence to us that this is pretty much teacher-centered. (Rater E 146:54; case 1) (my underlining)

Discussion

The main objectives of the present study were to investigate how CLA-HK experts assess Japanese EFL teachers' classroom English language proficiency using the CLA, to explore the rating differences between CLA-HK experts and IDS-J developers (whose benchmarks share the same basic concept), and to gain some insight into more effective use of the benchmark IDS and improvements that can be made to it.

The results of our study—which include the CLA-HK raters' assessments and, especially, the focus group interview—indicate that a classroom English benchmark should be able to help capture aspects of classroom English language proficiency that are indispensable for students' language acquisition. For example, an EMI teacher's output cannot create meaningful input for the students if it is not intelligible, but those EMI teachers who speak with some accent can be counted as acceptable so long as what they say is "internationally intelligible"—in other words, if it can be comprehended not only by native speakers and those who share the same mother tongue but also by non-native speakers too whose mother tongue is different. This is one of the standards in the area of *Pronunciation, Stress, and Intonation* that needs to be addressed in the revision of the IDS-J.

Another important point we have learned from the findings comes within the area *Interaction with Students*. It is that feedback must be of a kind that assists the students' output. Though praising helps bolster students' emotional confidence, its effect is likely to be limited and superficial unless the feedback addresses aspects of the students' linguistic competence. The revised IDS-J must be able to help assessors highlight this aspect, especially when it comes to giving feedback to teachers.

Furthermore, it should be noted that each rater has a personal view, as was seen in our discussion of "Which comes first, learner understanding or natural English?" Such discussion serves to unveil the multifaceted nature of classroom English proficiency—a feature that had not really been spotlighted before. Likewise, individual raters may also differ in their rigor. Herein lies the significance of having a benchmark that assists professional development, and the IDS has an enhanced potential for managing this.

Finally, the study confirmed that the target lesson should be a student-centered one so as to capture English language proficiency as conceptualized in benchmarks. Without appropriate data, it would be difficult for us (even as experts) to do the candidates justice in assessing the aspects of their language proficiency.

Conclusion

The focus of this study was not on a comparison of the assessment scores themselves as rated by the IDS developers and the CLA-HK assessors but was rather on the useful insights into a possible revision of the IDS-J and how it is used that could be gleaned by the exercise. Admittedly, there are some distinctive differences between the IDS-J and the CLA-HK with regard to (1) the type of lesson observed, with the IDS-J

requiring student-centeredness; (2) the rigor over errors, mere slips being distinguished from systematic error; and (3) pronunciation and accent. Such discrepancies confirm the need to make further improvement to the IDS in its content and its usage.

With these provisos in mind, we agreed that further improvements made to the IDS-J should be examined in light of the following points about its content and usage: (1) in the scheme, we need to ensure that teachers give lessons that include the five required elements; (2) we need to focus on "internationally acceptable" intelligibility, rather than on accent; (3) we should reflect the difference between systematic error and slips more clearly in the revision of the IDS-J (most probably from level 4 on) and, as assessors, we should be trained to discern the difference; and also (4) we need assessor training sessions in the use of the IDS-J to assist teachers' professional development. This was a point that came out of the assessors' critical analysis of the EMI performances and their insightful suggestions into IDS quality. As far as the content is concerned, having considered the two elements (2) and (3), we have come up with a revised version of the IDS-J (see Appendix). We have also reached the conclusion that we can maximize the potential of the IDS-J more fully by using it in the way suggested above ((1) and (4)).

Needless to say, all the insights are largely due to our reflection (as IDS-J developers) in this research project—including steps 1, 2, 3, and 4. As is often the case, and as was so with the LPATE for instance, the benchmark development inevitably requires a long, arduous, and continuing process of trial and experimentation. Nevertheless, the present study has made an indispensable step in the long journey toward meaningful contextualization of our benchmark IDS-J as a professional development tool.

Endnotes

[1]While a high-stakes test is used "to make important decisions about students, educators, schools, or districts for the accountability purpose," a low-stakes test is employed "to measure academic achievement, identify learning problems, or inform instructional adjustments, among other purposes." The difference between them lies not necessarily in their forms (how the tests are designed) but in their functions (how the results are used).(See Glossary of Education Reform) (https://www.edglossary.org/high-stakes-testing/)

[2]In the last 17 years (2001–2017), 11,026 teachers in Hong Kong have been assessed with CLA (see Assessment Reports of Language Assessment Proficiency for Teachers). (https://www.edb.gov.hk/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/lpat/lpat_assessment_reports.html)

[3]See The History of the Canadian Language Benchmarks (http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf)

[4]The four scales are *Integrated Diagnostic Scale*, *Reflective Analytic Scales*, *Function--Specific Scales*, and *Task-Specific Scales*. Nakata et al. (2012) briefly mentioned the first and the second elements, while Kimura et. al. (2017) further explained the first, the second, and the fourth ones.

[5]They are the research members of the KAKENHI (Grant-in-Aid for Scientific Research) project in Japan.

Appendix

Regarding the accuracy of classroom English proficiency (grammar, pronunciation), the possible revision of the IDS* suggested by the IDS developers are given below:

Grammar (Accuracy & Variety)

Level 4 (Proficient): Almost always uses an appropriate and wide variety of grammar in accordance with students' needs and proficiency flexibly.

➜ *Except for incidental minor mistakes*, almost always uses an *accurate* and wide variety of grammar in accordance with students' needs and proficiency flexibly.

Level 3 (Good): Mostly uses an appropriate and wide variety of grammar in accordance with students' needs and proficiency.

➜ *Except for incidental minor mistakes*, mostly uses an *accurate* and wide variety of grammar in accordance with students' needs and proficiency.

Level 2 (Acceptable): Generally uses grammar appropriately and reformulation (or self-correction) of occasional errors does not impede the flow of the lesson as a whole.

➜ *Generally uses grammar accurately* and reformulation (or self-correction) of occasional *mistakes* that does not impede the flow of the lesson as a whole.

**Level 1 (Not Yet Acceptable): Limited and inaccurate grammatical structure. Grammatical errors are often observed.

➜ *Mostly uses* limited and inaccurate grammatical structure. Grammatical errors are often observed.

Pronunciation (Accuracy & Naturalness)

Level 4 (Proficient): Almost always pronounces accurately and naturally, and also in a comprehensive way for almost all students by adjusting speed and pauses to them.

➜ Almost always pronounces accurately and naturally, *and expresses the intention of the communication with effective prosody. Also successfully conveys meaning to* almost all students by adjusting speed and pauses with sufficient flexibility to *their understanding*.

Level 3 (Good): Mostly pronounces accurately and naturally. Does not impede international communication. Also successfully conveys meaning to many students by adjusting speed and pauses to *their understanding*.

➜ Mostly pronounces accurately and naturally, *and expresses the intention of the communication with appropriate prosody*. Also successfully conveys meaning to many students by adjusting speed and pauses to *their understanding*.

Level 2 (Acceptable): Generally pronounces accurately (i.e., internationally intelligible) and reformulation (self-correction) of occasional mistakes (i.e., unacceptable ones as standard English) does not impede students' understanding.

➜ *Occasionally uses monotone prosody, but* generally pronounces accurately and reformulation of occasional mistakes does not impede students' understanding.

Level 1 (Not Yet Acceptable): Inaccurate and unnatural pronunciation confuses students.

➜ *No change*.

* The changed parts are italicized (See Kimura et al. (2017) for the details of the original version).

** Considering the nature of the benchmark IDS (i.e., not for assessment purpose but for professional development purpose), we decided to keep the description of Level 1 in each scale (which has a wider range than that in CLA) as they are.

Additional file

Additional file 1: Integrated Diagnostic Scale. (DOC 65 kb)

Abbreviations
CLA-HK: Classroom Language Assessment in Hong Kong; EDB: The Education Bureau; EMI: English medium instruction; ESP: English for Specific Purposes; HKEAA: The Hong Kong Examinations and Assessment Authority; IDS-J: Integrative Diagnostic Scale in Japan; LPATE: The Language Proficiency Assessment for Teachers; MEXT: The Japanese Ministry of Education, Culture, Sports, Science and Technology

Acknowledgements
We would like to offer our sincere appreciation to the two CLA experts for their ungrudging support and insightful suggestions in this research project. Thanks also should go to the research collaborators to our JSPS KAKENHI research projects.

Funding
This work has been supported by JSPS KAKENHI Grant Number 22530969, 26381199, and 17K04821.

Availability of data and materials
Interview data will not be shared because they will be used in other publications. Rating data (Table 2), the earlier version of the IDS, and the new version of the IDS are available.

Authors' contributions
YN, IO, YK, and NN discussed and carried out this study. YN drafted the manuscript, the other authors provided insightful comments, and SA as an expert of this field in particular. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]Faculty of Global Communications, Doshisha University, 1-3, Tatara Miyakodani, Kyotanabe, Kyoto 610-0394, Japan. [2]Faculty of Education, Ehime University, 3 Bunkyo-cho, Matsuyama, Ehime 790-8577, Japan. [3]Faculty of Medicine, University of Toyama, 2630 Sugitani, Toyama City, Toyama 930-0194, Japan. [4]International Education Center, Tokai University, 4-1-1 Kitakaname, Hiratsuka, Kanagawa 259-1292, Japan. [5]Faculty of Education, The University of Hong Kong, Hong Kong, China. [6]School of Education, University of New South Wales, Sydney, Australia.

Received: 22 May 2018 Accepted: 26 July 2018
Published online: 25 August 2018

References
ACTFL (2012). ACTFL Proficiency Guidelines 2012 (English). Retrieved from https://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012. Accessed 29 July 2018.

Alderson, JC, Brunfault, T, Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: insights from professional practice across diverse fields. *Applied Linguistics*, *36*, 236–260.

Andrews, S. (2007) Teacher language awareness. Cambridge: Cambridge University Press.

Brophy, J (2010). *Motivating students to learn*, (3rd ed.,). New York: Routledge.

Bruner, J (1983). *Child's talk: learning to use language*. New York: Norton.

Chapelle, CA, & Voss, E (2014). Evaluation of language tests through validation research. In AJ Kunnan (Ed.), *The companion to language assessment*, (1st ed., pp. 1081–1097). New York: Wiley.

Coniam, D, & Falvey, P. (1999). The English language benchmarking initiative: a validation study of the Classroom Language Assessment component. *Asia Pacific Journal of Language in Education*, *2*(2), 1–35.

Coniam, D, Falvey, P, Xiao, Y. (2017). An investigation of the impact on Hong Kong's English teaching profession of the language proficiency assessment for teachers of English (LPATE). *RELC Journal*, *48*, 115–133.

Dawson, C, Bodycott, P, Walker, A, Coniam, D. (2003). Continuing educational reform in Hong Kong: Issues of contextualization. *Educational Policy Analysis Archives*, *11*(5), 1–28 Retrieved from http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1416&context=coedu_pub. Accessed 29 July 2018.

Dearden, J (2014). *English as a medium of instruction – a growing global phenomenon: Phase 1 interim report*. London: British Council Retrieved from https://www.britishcouncil.org/sites/default/files/english_as_a_medium_of_instruction.pdf. Accessed 29 July 2018.

Douglas, D. (2001). Language for specific purposes assessment criteria: where do they come from? *Language Testing*, *18*, 171–185.

Education Bureau, The Government of the Hong Kong Special Administrative Region (EDB) (2017). Language Proficiency Assessment for Teachers (English Language) 2017 Assessment Report. Retrieved from http://www.edb.gov.hk/attachment/en/teacher/qualification-training-development/qualification/language-proficiency-requirement/2017LPAT_AssessmentReport_ENG.pdf. Accessed 29 July 2018.

Elder, C. (2001). Assessing the language proficiency of teachers: are there any border controls? *Language Testing*, *18*, 149–170.

Freeman, D. (2017). The case for teachers' classroom English proficiency. *RELC Journal*, *48*, 31–52.

Hong Kong Examinations and Assessment Authority (HEKAA) (2011). *LPATE handbook: the language proficiency assessment for teachers (English language)*. Hong Kong: Education Bureau Government of the Hong Kong Special Administrative Region.

Ikeno, O, Nakata, Y, Kimura, Y, Naganuma, N. (2016). Development of Task-Specific Scales for assessing Japanese teachers' use of English in English classes. *JACET Chugoku-Shikoku Chapter Research Bulletin, 13*, 171-183.

Kim, SHO, & Elder, C. (2008). Target language use in foreign language classrooms: practices and perceptions of two native speaker teachers in New Zealand. *Language, Culture and Curriculum*, *21*, 167–185.
Kimura, Y, Nakata, Y, Ikeno, O, Naganuma, N, Andrews, S. (2017). Developing classroom language assessment benchmarks for Japanese teachers of English as a foreign language. *Language Testing in Asia, 7*(3). https://doi.org/10.1186/s40468-017-0035-2.
Macaro, E. (2014). English medium instruction: time to start asking some difficult questions. *Modern English Teacher*, *24* Retrieved from https://www.modernenglishteacher.com/media/5377/macaro.pdf. Accessed 29 July 2018.
McKay, SL, & Brown, JD (2016). *Teaching and assessing EIL in local contexts around the world*. New York: Routledge.
McNamara, T. (2001). Language assessment as a social practice: challenges for research. *Language Testing*, *18*, 333–349.
MEXT (2010). Japanese MEXT's course of study guideline: high school (English). Retrieved from http://www.mext.go.jp/component/a_menu/education/micro_detail/__icsFiles/afieldfile/2010/01/29/1282000_9.pdf. Accessed 29 July 2018.
MEXT (2015). Regarding the result of "English Education Survey". Retrieved from http://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2016/04/05/1369254_3_1.pdf. Accessed 29 July 2018.
Nakata, Y. (2010). Improving the clssroom language proficiency of non-native teachers of English: What and how? *RELC Journal, 41*(1), 76–90. https://doi.org/10.1177/0033688210362617.
Nakata, Y. (2011). Teachers' readiness for promoting learner autonomy: A study of Japanese EFL school teachers. *Teaching and Teacher Education, 27*, 900–910.
Nakata, Y, Ikeno, O, Naganuma, N, Kimura, Y, Andrews, S. (2012). Classroom English language benchmarks for Japanese EFL teachers. *Proceedings of the JACET 51th international convension,* 20–27
Pasternak, M, & Bailey, KM (2004). Preparing nonnative and native English-speaking teachers: issues of professionalism and proficiency. In LD Kamhi-Stein (Ed.), *Learning and teaching from experience: perspectives on nonnative English-speaking professionals*, (pp. 155–175). Ann Arbor: The University of Michigan Press.
Sewell, A. (2013). Language testing and international intelligibility: a Hong Kong case study. *Language Assessment Quarterly*, *10*, 423–443.
Shintani, N (2016). *Input-based tasks in foreign language instruction for young learners*. Amsterdam: John Benjamins.
Sinclair, JM, & Coulthard, M (1975). *Towards an analysis of discourse: the English used by teachers and pupils*. London: Oxford University Press.
Tomasello, M (2003). *Construing a language: a usage-based theory of language acquisition*. Cambridge: Harvard University Press.
Tsui, ABM. (1985). Analyzing input and interaction in second language classrooms. *RELC Journal*, *16*, 1–30.
Tsui, ABM (1996). Reticence and anxiety in second language learning. In KM Bailey, D Nunan (Eds.), *Voices from the language classroom*, (pp. 145–167). Cambridge: Cambridge University Press.
Zimmerman, BJ (1998). Developing self-fulfilling cycles of academic regulation: an analysis of exemplary instructional models. In DH Shunk, BJ Zimmerman (Eds.), *Self-regulated learning: from teaching to self-reflective practice*, (pp. 1–19). New York: Guilford Press.