Gan et al. Language Testing in Asia (2017) 7:19
DOI 10.1186/s40468-017-0051-2

Language Testing in Asia

RESEARCH Open Access

ESL students' oral performance in English language school-based assessment: results of an empirical study

Zhengdong Gan[1*], Emily Pey Tee Oon[1] and Chris Davison[2]

* Correspondence: zhengdonggan@umac.mo
[1]Faculty of Education, University of Macau, Macao, People's Republic of China
Full list of author information is available at the end of the article

Abstract

Background: The English language school-based assessment (SBA) component of the Hong Kong Diploma of Secondary Education (HKDSE) Examination is innovative in that the assessment tasks involve assessing English oral language skills in a high-stakes context but they are designed and implemented in the ESL classroom by school teachers in light of a regular reading and viewing program or the elective modules integrated into the school curriculum. While this certainly is a positive move towards better congruence between the teaching, learning, and assessment activities, there has been concern whether the teachers are capable of applying the assessment criteria and standards consistently in spite of going through a variety of standardization meetings and sharing discussions initiated and mandated by the Hong Kong Examination and Assessment Authority (HKEAA). In other words, there has been concern about the extent to which results provided from teachers in different schools are comparable. Also, how may task difficulty be reflected in students' assessment results across the two SBA task types? It was to provide some research evidence on matters relating to these issues associated with teacher assessment results that the study described here was carried out.

Methods: The study, with the help of Rasch analysis, aims to examine the psychometric qualities of this English language school-based assessment, how students' assessment results may vary across different schools, and how task difficulty may vary across the two different task types.

Results: The findings indicated the following: (1) among the three schools involved in this study, two band 2 schools demonstrated similar abilities across all task domains as there were no significant differences in students' SBA results in all assessment domains between these two band 2 schools. Significant differences were found in some assessment domains between the two band 2 schools and the band 3 school; (2) an obviously more fine-grained pattern of difference in difficulty levels of different assessment domains was observed in students' assessment results across the two task types in this study than in previous studies.

Conclusions: Implications of the results for teacher assessor training and test task development are discussed.

Keywords: School-based assessment, Oral performance, Rasch analysis

Background

In contrast to large-scale standardized testing in which the assessor is usually someone who must remain objective and uninvolved throughout the whole assessment process, school-based assessment tends to be embedded in the regular curriculum and assessed by a teacher who is familiar with the student's work (Davison 2007). Davison maintains that school-based assessment derives its validity from building into its actual design the capacity for triangulation and the collection of multiple sources and types of evidence under naturalistic conditions over a lengthy period of time. Consequently, "the reliability of the assessment was also enhanced by having a series of assessments (rather than just one) by a teacher who was familiar with the student and by encouraging multiple opportunities for assessor reflection and standardization." (Davison 2007, p. 51). In other words, teachers are in the best position to determine the quality of student achievement over time and at specific points and to improve student learning (Wyatt-Smith et al. 2010).

However, drawing on her qualitative observation, Sadler (1998, p. 80–82) made explicit the typical intellectual and experiential resources teachers rely on when making a judgment in classroom assessment:

- Superior knowledge about the content or substance of what is to be learned
- Deep knowledge of criteria and standards [or performance expectations] appropriate to the assessment task
- Evaluative skill or expertise in having made judgments about students' efforts on similar tasks in the past
- A set of attitudes or dispositions towards teaching, as an activity, and towards learners, including their own ability to empathize with students who are learning; their desire to help students develop, improve, and do better; their personal concern for the feedback and veracity of their own judgments; and their patterns in offering help

Implicit in Sadler's observation is thus that teacher judgments might be characterized as remaining responsive to the influence of other knowledge and skills rather than the stated standards and criteria. Clapham (2000) further commented:

> A problem with methods of alternative assessment, however, lies with their validity and reliability: tasks are often not tried out to see whether they produce the desired linguistic information; marking criteria are not investigated to see whether they 'work'; and raters are often not trained to give consistent marks. (p. 152).

In a survey of a high-profile school-based assessment initiative in Hong Kong (Davison et al. 2010), teacher comments such as "I would like the HKEAA to take up my marks to see if I have interpreted the criteria correctly" revealed a lack of confidence among teachers about this teacher-mediated and context-dependent assessment initiative, with many doubting that they had the required knowledge and skills to carry out the assessment properly. Although this English language school-based assessment component has been implemented in schools for nearly 10 years, there has been almost no empirical evidence to illustrate the extent to which teacher assessment results from one school are comparable to results of another school. Also, to what extent does difficulty level of different task domains vary across the two task types in this assessment? It was

to provide some research evidence on matters relating to teacher assessment results that the study described here was carried out.

School-based English language assessment (SBA) scheme in Hong Kong

The literature on school-based assessment has been growing for more than two decades (Davison 2007; Meisels et al. 2001; Gan 2012; Gan 2013; Qian 2014). School-based assessment, as an alternative to testing, in the form of greater use of teachers' assessment of their own students, has become increasingly popular in many countries over the world. Such curriculum-embedded performance assessments often defined as integrated parts of students' learning experience rely heavily on teacher judgment. They differ from external assessments in that curriculum-embedded performance assessments are integrated into the daily curriculum and instructional activities of a classroom (Meisels et al. 2001). The thinking behind the curriculum-embedded performance assessments is based on a social-constructivist view of learning (Vygotsky 1978). The use of this curriculum-embedded performance assessment is often advocated on the grounds that it can be conducted as part of teaching and so provide formative feedback to students, thus improving their learning (Crooks 1988). What characterizes this type of curriculum-embedded performance assessment is that both the teacher and students are actively engaged with every stage of the assessment process in order that they truly understand the requirements of the process, and the criteria and standards being applied (Price et al. 2007). Essential to the operation of this type of assessment is the teacher's ability to reconcile the dual role that they are required to take in both promoting and judging learning (Harlen 2005). Harlen points out that the task of helping teachers take up this dual role can be particularly difficult in countries where a great deal of emphasis is given to examinations results. For example, Choi (1999) suggested that in a highly competitive examination-driven school system such as Hong Kong's, success of a school-based assessment initiative hinges on assessment training and resource support provided for teachers. Choi, however, mentioned another difficulty in introducing a school-based assessment initiative is to ensure credibility for school-based assessment. This means that an effective and efficient quality assurance and quality control system needs to be established so that the users of examination results can be assured of the reliability of this scheme of assessment and have confidence in the teachers' judgments.

The school-based assessment (SBA) scheme in Hong Kong started out as a component of the Hong Kong Certificate of Education Examination (HKCEE) English Language in 2006. This assessment scheme which was collaboratively initiated by the Education Bureau (EDB) and the Hong Kong Examinations and Assessment Authority (HKEAA) is innovative in that assessments are administered in schools and marked by teachers in the context of public assessment. Grounded within an "assessment for learning" framework, it is now incorporated into the new Hong Kong Diploma of Secondary Education (HKDSE) English Language Examination, adopting a standards-referenced assessment system, aiming to not just report on the full range of educational achievement but also motivate learning in Hong Kong secondary schools. In addition to the fact that this assessment scheme accounts for 15% of the total subject mark in the HKDSE, this SBA component seeks to provide a more comprehensive evaluation of learners' achievement by assessing those learning objectives which can hardly be assessed in public assessments while concurrently enhancing the capability for student

self-assessment and life-long learning (Davison 2007). Given the nature of multiple functions of this SBA component, we believe this current school-based English language assessment can best be defined as:

> The process by which teachers gather evidence in a planned and systematic way in order to draw inferences about their students' learning, based on their professional judgment, and to report at a particular time on their students' achievements (Harlen 2005, p. 247).

According to HKEAA, these two kinds of assessment tasks build on two different kinds of learning programs embedded in the school curriculum in Hong Kong. One is a reading/viewing program in which students read/view four texts over the course of 3 years and undertake an individual presentation or a group interaction based on the books/videos/films that they have read/viewed. The other is the elective module(s) in the school curriculum where students carry out an individual presentation or a group interaction based on the knowledge, skills, and experience gained in these elective modules.

Although SBA underwent a detailed research, development, and distribution process and bears the advantages of providing teachers with a formative view of the progress of individual students and allowing them to address more effectively the specific needs of their students (Yip and Cheung 2005; Carless and Harfitt 2013), challenges and controversy arose particularly when assessment for both formative and summative purposes is integrated into the regular teaching and learning process, with school teachers involved at all stages of the assessment cycle, from planning the assessment program, to identifying and/or developing appropriate assessment tasks right through to making the final judgments (Davison 2007). While responses of teachers and students to the underlying philosophy of SBA and its emphasis on improving the quality of teaching and learning were generally very positive, concern about the comparability of SBA scores across schools has been pervasive and still continues, with some more experienced teachers being even more vocal with regard to negative comments towards the administration of SBA in the initial stage (Qian 2014).

Reliability is often defined as the consistency of measurement (Bachman and Palmer 1996). In other words, the reliability of a test or assessment has to do with the consistency of scoring and the accuracy of the administration procedures of the test or assessment (Chiedu and Omenogor 2014). Chiedu and Omenogor suggest that in the case of teacher-directed classroom assessment, two teacher assessors may not necessarily interpret the assessment criteria the same way. In addition, as teacher-directed classroom assessment may vary in different contexts at different times, it may lead to inconsistent assessor judgment (McNamara 1996). It has thus been widely believed that a major source of unreliability is the scoring of a test or assessment. Undoubtedly, reliability is as an important issue for school-based assessment as for traditional testing. Currently, in the case of English language SBA in Hong Kong, the following methods, *within-school standardization*, *inter-school sharing*, and *HKEAA's statistical moderation*, are adopted by the HKEAA (2016, p. 22) to ensure the reliability and consistency of SBA scores across schools. Below is a description of each of these four assessment training methods.

Within-school standardization

"Within-school standardization" means that if there is more than one subject teacher teaching the subject to the same cohort of students in the school, it is necessary for the

teachers involved to agree on the criteria for awarding marks so that the same standard of assessment is applied to all students. Specifically, teachers teaching the same cohort of students bring samples of video-recorded assessments of different levels (e.g., the three highest and the three lowest assessments) to the school-level standardization meeting where the video-recorded assessments are shown and discussed. The discussions at this school-level standardization meeting may lead to adjustments to scores across classes in the school. This school-level standardization ensures that all the teachers involved in SBA in the school will achieve a clear understanding of the shared expectations of what students at particular levels should be able to do in order to achieve a certain score.

Inter-school sharing

Following the within-school standardization meeting, "Inter-school sharing" meetings are organized by SBA District Coordinators. At the end of the school year, the SBA District Coordinator will organize an inter-school meeting for professional sharing among the schools within the group. The School Coordinators bring samples of video-recordings and assessment records to this inter-school meeting where these samples of student performance from different schools will be viewed and discussed with reference to the assessment criteria. Each School Coordinator needs to report back to colleagues in their own schools. If it is apparent that a particular school's scores are markedly higher or lower as a whole than those from the other schools as a whole, the school team may wish to review their scores.

HKEAA's statistical moderation

Despite the school-level teachers' participatory and reflective professional sharing in the implementation of SBA, there is still the likelihood that teachers in one school may be harsher or more lenient in their judgments than teachers in other schools. Given this concern, a statistical moderation method is adopted by HKEAA in moderating the SBA assessments submitted by schools, with the aim to ensuring the comparability of SBA scores across schools. This statistical moderation is done by adjusting the average and the spread of SBA scores of students in a given school with reference to the public examination scores of the same group of students, supplemented with review of samples of students' work. The statistical moderation results will be compared to the results from the sample review. Potential adjustments will be made to the statistical moderation results so that the final moderated scores of these schools can properly reflect the performance of their students in the SBA.

Kane (2010) makes a distinction between procedural fairness and substantive fairness. Procedural fairness can be said to require that all test takers take the same test or equivalent tests, under the same conditions or equivalent conditions, and that their performances be evaluated using the same rules and procedures. Substantive fairness in testing requires that the score interpretation and any test-based decision rule be reasonable and appropriate and 'that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership'(AERA et al. 1999, p. 74). In other words, substantive fairness is concerned with how well the program functions for different groups, and it requires that scores have comparable meaning in different groups. While the above

school-level processes of systematic, participatory and reflective professional sharing may indeed be helpful in mitigating stakeholders' potential concern about fairness of SBA scores across schools, there has been almost no empirical evidence to illustrate the extent to which SBA results across different types of schools are comparable. To fill in this research gap, the present study, with the help of Rasch analysis, aims to examine how students participating in SBA in different schools may vary with regard to SBA scores in the assessment tasks.

Task-based L2 performance in the research literature

Currently, there are two competing theoretical perspectives on task-based L2 performance aiming to account for the impact of task type and task conditions on L2 spoken performance, the Tradeoff Hypothesis (Skehan 2009; Skehan 2014) and the Cognition Hypothesis (Robinson et al. 2009, Robinson 2007). Skehan's theoretical framework views limitations in attention as fundamental to second language speech performance, which entails a need to analyze what attentional and working-memory demands a task makes and the consequences this may have for different language performance dimensions often referred to as accuracy, fluency, and complexity. Consequently, it is often assumed that more demanding tasks are likely to result in prioritization of fluency over accuracy and complexity and that tasks based on familiar or concrete information favor a concern for accuracy. Also, within this Tradeoff Hypothesis, it is suggested that interactive tasks or tasks requiring transformation or manipulation of materials or tasks which have had pre-task planning are likely to lead to greater linguistic complexity. Standing in clear opposition to Skehan's Tradeoff Hypothesis, Robinson's (2009, 2007) Cognition Hypothesis claims that there is no limit to human attentional resources and as such human mind can attend to different aspects of performance if certain conditions are met and that language learners can access multiple attentional pools that do not compete and depletion of attention in one pool has no effect on the amount remaining in another. Robinson (2007) also argues that the more demanding a task is in terms of its content, the more complex and accurate its linguistic performance will be.

Empirical studies that were guided by either Skehan's or Robinson's framework and conducted in pedagogic con-texts, however, yielded mixed results. For example, Bygate (1999) examined the complexity of the language of Hungarian secondary EFL learners on a monologic narrative task and an argumentation task and found that the narrative tasks might stretch the speakers more in terms of complexity of syntactic and lexical processing. Bygate's study finding appeared to be echoed in the Michel et al. (2007) study which revealed that the dialogic (i.e., interactive) task tended to elicit shorter and structurally simpler sentences than the monologic narrative task, although Michel et al. also found that students made significantly fewer errors and were significantly more fluent in the dialogic task condition. In other words, Michel et al.'s study suggests that interactivity may affect structural complexity negatively. It was thus apparent that Skehan and his colleagues' (Foster and Skehan 1996; Skehan and Foster 1997) observation that more interactive tasks lead to more complex language performance did not find support in the Bygate and Michel et al. (2007) studies. In language testing contexts, a few studies (e.g., Fulcher 1996; Bachman et al. 1995) reported significant but small differences in test scores across different types of test tasks. More recently, a number of studies conducted in experimental language testing settings that replicated Skehan's or Robinson's framework

concerning the impact of task performance conditions on task performance revealed results that did not lend much support to either of the their theoretical frameworks. Given the mixed results of these studies on the relationship between task type and task performance, it is clear that this issue warrants further empirical research.

The context for the present study is innovative in that the assessment tasks in this study involve speaking in a high-stakes language assessment context but they are designed and implemented in the ESL classroom by school teachers in light of a regular reading and viewing program or the elective modules integrated into the school curriculum (Gan 2013). The processes of selecting appropriate assessment tasks and making the actual assessments are undertaken collaboratively among teachers concerned, taking into account the students' backgrounds, needs, and their skills. All the teachers involved in the assessment, however, need to go through a series of within-school and inter-school standardization meetings and discussions organized by the HKEAA to help them to develop a clear understanding of the shared expectations of what students at particular levels of performance should be able to do to achieve a certain grade.

Building on the research discussed above, the present study focuses on the following research questions:

1. What is the variation of SBA results across schools in Hong Kong?
2. How may task difficulty be reflected in students' assessment results across the two SBA task types?

Methods

Participants

The study is part of a large-scale longitudinal project of investigating teachers and students' perceptions of a high-profile school-based assessment initiative in Hong Kong and using various measures to validate assessment tasks and assessment results. In an earlier related study, a convenience sample of 373 secondary Form 6 students from three different schools completed a questionnaire about their perceived difficulty of the two task types on the school-based assessment. The students also reported on their assessment results from the two assessment tasks. The study reported in this paper focused on analysis of the students' assessment results collected in the earlier study. Among the three schools involved in the study, schools A and B are both catholic schools where English is used as the medium to teach core subjects such as English, Math, Chemistry, and Physics. School C became a Chinese-medium school after 1997, and at the time of this study, school C was making efforts to build up better discipline and learning atmosphere among the students. Note that schools A and B are ranked as band 2 school whereas school C is ranked as band 3 school in the traditional local school rankings.

Procedures

Prior to students' participation in the questionnaire survey, students' performance in the two SBA tasks were assessed by their teachers who followed the assessment criteria for both group discussion and individual presentation that cover six levels (level 1 represents the lowest level, and level 6 represents the highest level) of oral English proficiency in the four major domains of English language performance. The two task types are defined by HKEAA (2016) as follows:

An *individual presentation*, which may be quite informal, is defined as a single piece of oral text in which an individual speaker presents some ideas or information over a sustained period (3–5 min), with the expectation that they will not be interrupted. An individual presentation requires comparatively long turns, hence generally needing more pre-planning and a more explicit structure to ensure coherence. A presentation may be followed by questions or comments from the audience, but this exchange is not mandatory for the assessment for the individual presentation.

A *group interaction* is defined as an exchange of short turns or dialog with more than one speaker on a common topic. An interaction is jointly constructed by two or more speakers, hence generally needing less explicit structuring but more attention to turn-taking skills and more planning of how to initiate, maintain, and/or control the interaction by making suggestions, asking for clarification, supporting and/or developing each other's views, and disagreeing and/or offering alternatives.

In each of the individual presentations or group discussions, each participant thus received a separate score for each of the four domains of assessment criteria, as well as a global score as a result of the aggregation of the domain scores (Gan 2012).

Data analysis

In some of the previous test, validation studies, test psychometric properties, and result interpretations were analyzed typically through conventional analysis methods. For instances, internal consistency of test items in the form of Cronbach's alpha is usually examined for the indication of reliability; face and content validity are obtained solely from a panel of experts; raw scores from each item were summed across for a total mean score for comparison of students' performance or for parametric statistical test examination. Such conventional analyses of raw scores assumed interval-scale data for an ordinal-scale data (Wright 1999) where parametric statistical tests are not readily to be performed on (Wright and Master 1982; Boone et al. 2014; Liu 2010). When parametric test was done on ordinal data, the results have an element of error. In other words, the reliability and validity of data are jeopardized. In the present study, the psychometric properties of the test were assessed by Rasch modeling analysis and raw scores were transformed into interval data (Rasch estimates in unit *logit*) for the conduct of parametric statistical test—these features clearly advanced the precursory studies.

In the current paper, the school-based English language assessment scores from 373 secondary Form 6 students from three schools were analyzed using Rasch analysis (Rasch 1980) with FACETS software (Linacre 2017). In the analysis, each separate domain of task performance of the two SBA assessment tasks is referred to as an assessment "item," scored on a 6-point scale (see Appendixes 1 and 2). A total of eight assessment items were included for analysis. This enables the psychometric quality of the instrument to be assessed. For this purpose, principal component analysis of residuals, fit statistics, and Rasch separation indices were examined. Rasch model was used to transform the raw scores into interval-scale data for analyses. Specifically, raw scores were transformed into Rasch's estimates that are linear and readily used for conventional statistical analyses, e.g., ANOVA for variables comparisons. In order to evaluate whether scores on the eight items of the two assessment tasks were significantly different across schools, interaction analysis between item difficulty and schools was conducted. In order to examine the relative task difficulty across the two SBA task types, the difficulty estimates of the eight items were compared.

Results

Psychometric features of the assessment

Rasch model expects unidimensionality where scores are measures of only one latent trait. While principal component analysis (PCA) of residuals identifies potential secondary dimension that distorts the measurement of the latent trait (Linacre 2014), it assists unidimensionality assessment through variance explained by Rasch's modeling measures. The data is assumed to be unidimensional if the variance explained by Rasch measures is greater than or equal to 50% (Linacre 2014). For the present study, the PCA of residual test reporting 77.5% of variance was explained by Rasch measures. This is an indication that the data are sufficiently unidimensional and appropriate for Rasch analysis—an attribute of construct validity (Bond and Fox 2015) and of strong evidence that the scores are interpretable.

Fit statistics are also indicators for unidimensionality. The fit statistics is assessed through *Infit* and *Outfit* Mean Squares (MnSq). Infit MnSq are derived from on-target performance scores while Outfit MnSq are influenced more by the off-target scores. Data that fit the Rasch model perfectly will yield a fit of 1. This ideal situation is impossible in real world from actual data. A MnSq fit range between 0.60 and 1.40 indicated good adhesion to the model (Bond and Fox 2015; Wright and Linacre 1994). Misfitting statistics indicated that test items may measure more than one latent trait. Results to the items staying outside the acceptable range should be interpreted with caution. Table 1 shows that all items reported acceptable Infit and Outfit MnSq with values ranging between 0.86 and 1.18. In addition to the PCA results reported earlier, the item fit statistics indicated that the data were unidimensional and that item performed according the Rasch model's expectations.

Rasch modeling two separation indices providing information on whether the person and item estimates estimated by Rasch model are reliable. Person separation index indicates replicability of person ordering while item separation index indicates replicability of item placement on an interval scale (Bond and Fox 2015). The widely accepted threshold for the separation index is 3.0 (Bond and Fox 2015). Person and item separation indices for the present study were 5.07 and 4.31 (corresponding to 7.09 person strata and 6.08 item strata)—these results indicate that this sample and items are

Table 1 Item fit statistics

Entry	Name	Measure	SE	Infit MnSq	Outfit MnSq
1	disProdel	0.06	0.08	1.01	1.05
2	disComstr	− 0.12	0.08	1.16	1.18
3	disVoclan	0.37	0.08	0.86	0.86
4	disIdeorg	− 0.67	0.08	0.91	0.86
5	indpreProdel	0.19	0.08	0.87	0.88
6	indpreComstr	0.33	0.08	1.14	1.11
7	indpreVoclan	0.25	0.08	0.86	0.88
8	indpreIdeorg	− 0.42	0.08	1.10	1.16

Notes: Entries 1–4 refer to the four performance domains (i.e., pronunciation and delivery, communication strategies, vocabulary and language patterns, ideas and organization) of the SBA group interaction task; entries 5–8 refer to the four performance domains (i.e., pronunciation and delivery, communication strategies, vocabulary and language patterns, ideas and organization) of the SBA individual presentation task. Each domain of either task is scored on a 6-point scale. See Appendixes 1 and 2

separable into 6–7 levels of ability or difficulty levels, respectively (Bonk and Ockey 2003), from which they indicated that the person and item estimates from the Rasch analysis are reliable and are replicable on interval scales (Bond and Fox 2015).

What is the variation of SBA results across different schools in Hong Kong?

An interaction analysis between item difficulty and schools was conducted to examine the difference of students' oral English performance on the eight assessment items across the three different schools. The result of chi-square test showed that the interaction effect was significant [χ^2 = 111.3, p < .05]. In other words, generally, students from different schools demonstrated significantly different performance on the items.

Rasch estimates are indicators of students' ability and item difficulty. A positive value indicates higher ability and higher difficulty; in contrary, a negative value indicates lower ability and lower difficulty. Students from school C scored higher on discussion task (Fig. 1) as they reported lower Rasch-calibrated item difficulties across the four group discussion task domains (items). Item 1 (disProdel) (– 0.34 logit) and item 2 (disComstr) (– 0.41 logit) were particularly easier for student from school C than they were for students from schools A and B. The item difficulty of item 1 for schools A and B were 0.34 logit and 0.42 logit, while the item difficulty of item 2 were 0.18 logit and 0.07 logit, respectively. The differences between schools A/B and C were significant (p < .05). The differences of item difficulty of item 3 (disVoclan) and item 4 (disIdeorg) between students from school B and school C were also significant (p < .05). It is obvious that students from schools A and B demonstrated similar abilities across the discussion items, as no significant difference on item difficulty was observed between them on all discussion task domains.

In general, students from school C showed poorer performance on individual presentation task domains, especially on items 6, 7, and 8. The item difficulty of these three items for students from school C were 0.64 logit, 0.77 logit, and – 0.03 logit

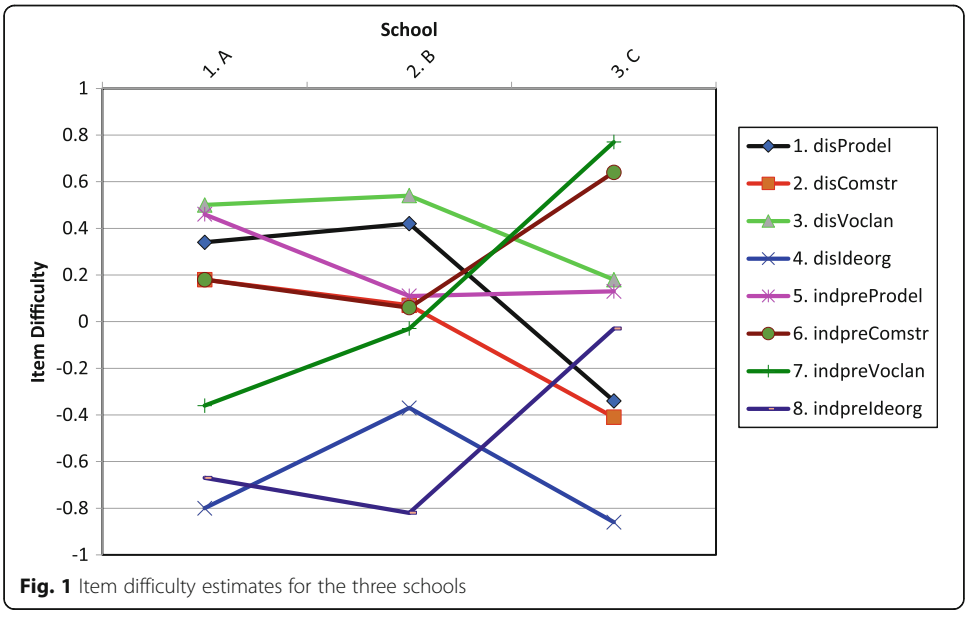

Fig. 1 Item difficulty estimates for the three schools

respectively. In contrary, students from schools A and B performed significantly better on these items ($p < .05$). The item difficulty of these three items were 0.18 logit, – 0.36 logit, and – 0.67 logit for students from school A and 0.06 logit, – 0.03 logit, and – 0.82 logit for students from school B. Students from schools A and B demonstrated similar performance across the individual presentation task items without any significant difference on item difficulty observed.

The pattern of relative item difficulty for students from different schools is clearer in Fig. 2. Students from schools A and B demonstrated similar performance pattern. The items in discussion task appeared more difficult for school A/B than they were for school C. In contrast, items in individual presentation task were easier for school A/B than they were for school C.

How may difficulty of items (i.e., assessment domains) be reflected in students' assessment results across the two SBA task types?

Figure 3 lays out the locations of the students and the items on an interval scale. The first column is the logit scale, and the second and third columns graphically described the locations of the students and the eight items, respectively. The fourth column is the rating scale of the items. This map transformed the student scores and item scores on a common interval scale in logit unit. For the present study, the logit scale runs from – 10 to + 9 logits. Students towards the top of the figure were higher in ability than students staying at the bottom. Items near the top are more difficult items while those near the bottom are less difficult items.

Across the two task types, item 4 (dislde org) and item 8 (indpreldeorg) appeared to be the easiest items to students (Fig. 3); the former is a group discussion item while the latter an individual presentation item. Item 6 (indpreComstr), item 3 (disVoclan), and item 7 (indpreVoclan) emerged as the most difficult items (Fig. 3); item 6 and item 7 are individual presentation items while item 3 is a group discussion item. The remaining items 2 (disComstr), 5 (disProdel), and 1 (disProdel) appeared to be of

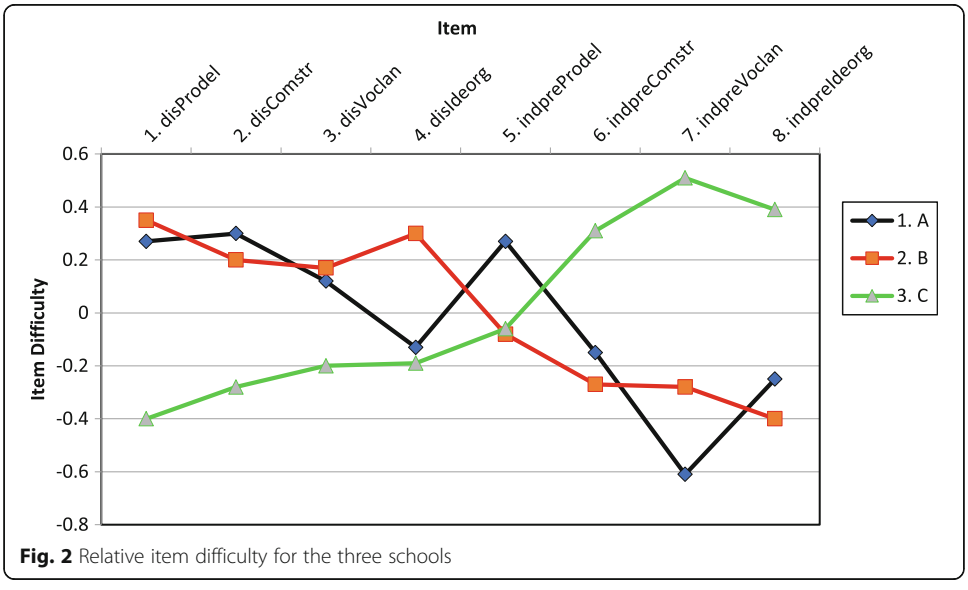

Fig. 2 Relative item difficulty for the three schools

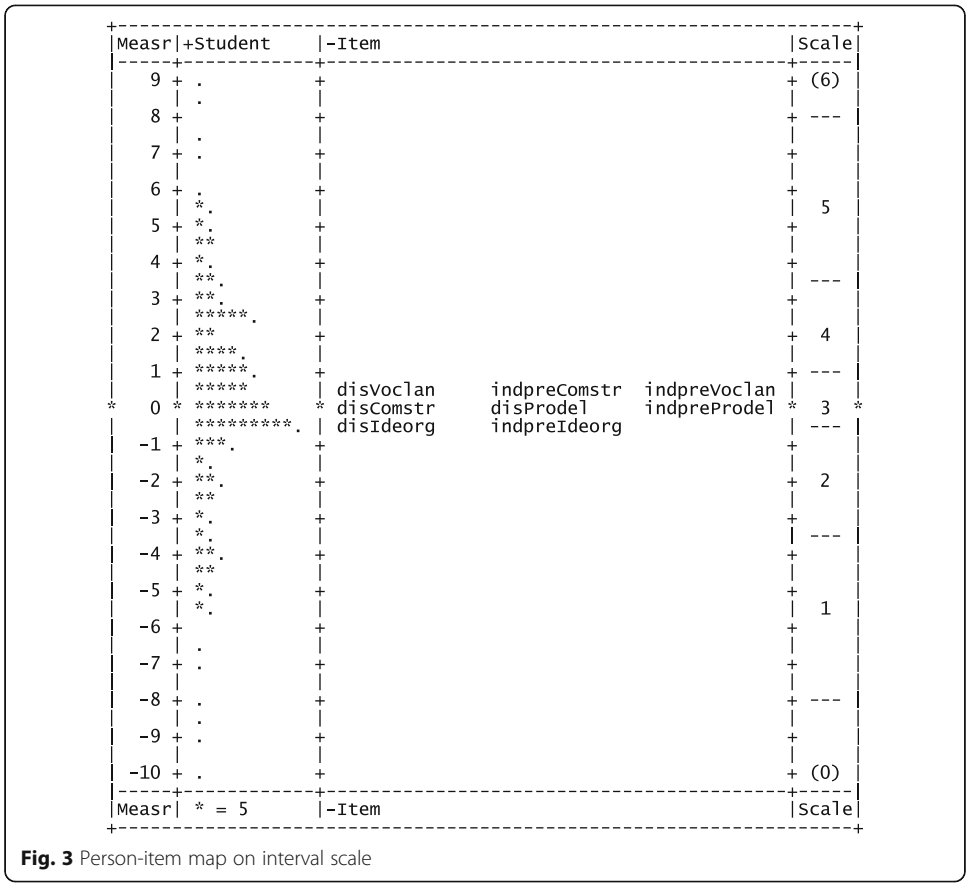

Fig. 3 Person-item map on interval scale

medium difficulty level relatively. These results suggest some more fine-grained pattern of difference in item difficulty across the two SBA task types, which was not reported in previous studies.

Discussion

Psychometric qualities of the English language school-based assessment

In the English language school-based assessment, the teacher assessor, who has received rater training organized by the HKEAA before undertaking assessment, sits nearby, assesses each participant, and assigns scores to students. Each student thus receives two independent ratings for their oral English performance in either the individual presentation or group interaction task and is scored on pronunciation and delivery, communication strategies, vocabulary and language patterns, and ideas and organization. Raw scores for each of the assessment tasks were assigned on a scale of 0–6 on each of four rating categories, for a total score of 24. In conducting data analysis of test datasets, the assumption of unidimensionality is perhaps one of the most-discussed features of Rasch models (Bonk and Ockey 2003). In our study, the statistics reported above display adequate psychometric unidimensionality, suggesting the English language school-based assessment tends to assess a unidimensional latent trait, i.e., the oral English proficiency, as represented by ratings on four scoring categories and thus providing evidence of construct validity. The statistics also show that the

items in the SBA had satisfactory fit statistics, indicating that all items performed in a consistent way as expected. The person and item separation indices shown in the summary statistics are above the widely accepted threshold for the separation index, indicating that the SBA can differentiate levels of proficiency. This means that the Rasch model generated in our analysis could reliably separate examinees by ability.

Bonk and Ockey (2003) used Rasch measurement in their study of a group oral assessment in English language at a Japanese university, in which groups of three or four students were assessed in conversation. Examinees were scored on five soring categories, i.e., pronunciation, fluency, grammar, vocabulary/content, and communication skills/strategies. Although items in the form of these five scoring categories were found to show acceptable fit, "communication skills" and "pronunciation" were the categories with a possible degree of misfit. Two scoring categories of our study (see Appendixes 1 and 2) also measure "communication skills" and "pronunciation," but unlike Bonk and Ockey's study, these two categories as well as the other two categories in our study all demonstrate good fit. This means that all these assessment categories (pronunciation and delivery, communication strategies, vocabulary and language patterns, ideas and organization) obviously belong to the same measurement domain. This makes sense as the focus of the school-based assessment is on the students' speaking ability to discuss issues in depth and to convey their ideas clearly and concisely rather than memorization skills or their ability to provide highly specific factual details about what they have read or viewed.

Variation of SBA results across schools in Hong Kong

This study showed that students from school C demonstrated significantly poor performance on three assessment domains (i.e., communication strategies, vocabulary and language patterns, and ideas and organization) in the individual presentation task compared with school A and school B. However, somewhat unexpectedly, students from school C scored significantly higher on two assessment domains (i.e., pronunciation and delivery, and communication and strategy) in the group discussion task than students from school A or school B, given the fact that school C is a government-funded band 3 school. At the time of this study, school C was struggling hard to improve its teaching quality and discipline maintenance among the students. There are two possible interpretations of school C's higher performance on those two assessment domains. First, as a typical practice in many government-funded band 3 schools in Hong Kong, these schools tend to designate a couple of classes from each grade as "elite classes." Such "elite" classes usually have the privilege of access to the best teaching and learning resources in the school. For example, these classes are usually taught by the best English teachers in the school and may also participate in extra-curricular English learning tutorials offered by native-English speaking teachers in the school. In this study, there was the likelihood that a considerable proportion of elite class students from school C might have participated in this study. Second, there was the possibility that some teacher assessors from school C might have been lenient in assessing their students' oral performance in some assessment domains in the group discussion task in the SBA. Overall, this study indicates that students from school C in this study

were likely to demonstrate relatively unstable language performance in the English language SBA in Hong Kong. This implies that in spite of the school-level teachers' participatory sharing in the SBA standardization processes, there is still the likelihood that there is variance in being harsh or lenient in their judgments of students' performance in different assessment domains in school C. This study thus points to the need for HKEAA to adopt a statistical moderation method in moderating the SBA assessment results submitted by schools to ensure comparability of SBA scores across schools.

Two of the three schools involved in this study, schools A and B, are catholic direct-subsidy schools that use English as the medium to teach the core school subjects, and have been ranked as band 2 schools in Hong Kong. This study shows that students from these two catholic subsidy schools demonstrated no statistically significant differences in their school-based English language assessment results, suggesting that teachers' assessment scores appeared to be comparable across these two schools. In other words, teacher judgments of the student performance from these two band 2 schools on the two English language SBA tasks tend to be consistent. Such potentially reliable judgment of students' performance on the SBA might have to do with a range of standardization procedures within or across schools that enable teachers to meet together, look at/listen to/discuss student oral samples, the tasks students have done, and talk about why they think a sample is at a level on each domain. These procedures thus likely constitute the important processes that contribute to understanding and to common grounds among English teachers involved in the SBA.

Difference in difficulty of different assessment domains across the two SBA task types

The notion of "task-induced variation" (Ellis 1994) means that a particular type of task that a learner is asked to perform will result in variation (Rahimpour 2007). This is echoed by Tarone (1990) who argues that as second language learners perform different tasks, their production of some grammatical, morphological, and phonological forms will vary in a particular manner. Gan (2013) examined how learner L2 oral performance may vary across two different task types in the current school-based assessment in Hong Kong by analyzing both the discourse produced from the tasks and the teacher rater assessments of students' task performance. Gan's study revealed a general trend towards higher assessment scores on most of the assessment domains in individual presentation task than in the group discussion task. It needs to be pointed out that only 30 students' assessment performance from one particular secondary school in Hong Kong was analyzed in the Gan study. With the help of Rasch analysis, the present study examined the teacher rater assessments of 373 students across three different schools and revealed an obviously more fine-grained pattern of difference in difficulty levels of different assessment domains observed in students' assessment performance across the two SBA task types. Item 4 (disldeorg) of the group discussion task and item 8 (indpreldeorg) of the individual presentation task appeared to be the easiest task domains to students. This result could be associated with the possibility that while assessing student oral performance, the teacher rater was likely to attend more to the grammatical, lexical, or phonological features of the test candidate's language use than to

organization of their ideas. This appears to corroborate the result that item 3 (disVoclan) of group discussion task and item 7 (indpreVoclan) of individual presentation task emerged as the most difficult items as these items represent performance domains on which the teacher assessor was more likely to base their decisions (Gan 2012). Note that item 6 (indpreComstr) was also one of the most difficult items. This might be due to the possibility that the condition under which the learner performed individual presentation task resulted in the learner concentrating on accuracy and fluency of their language production but overlooking use of interactional skills. Consequently, these results show that different aspects of the two SBA tasks may have different strengths in measuring students' speaking proficiency in the school-based assessment context. In other words, the result provides evidence that the two SBA task types could be used to complement each other in measuring the same construct of oral language proficiency as they claim to measure. In the past decades, there has been anxiety among educators and researchers about the reliability of the group oral discussion format in the testing literature. The results of this study lead us to concur with Bonk and Ockey that the group oral may also be a reasonably solid basis upon which to make a valid overall decision about students' L2 oral ability.

Conclusions

This study was motivated by the concern in both research and practice that teachers from different schools might not be able to provide comparable results, given teachers' necessarily subjective judgments and interpretations of assessment data. We were thus interested to examine the extent to which teachers' assessment results from three different schools were comparable. The results suggest that assessment results from two band 2 schools appeared generally comparable as there was no significant difference in students' SBA results in most assessment domains across the two schools. Teachers' assessment scores of students from the band 3 school in this study could be less stable occasionally as students from this school scored significantly lower on some assessment domains but significantly higher on some other domains compared with the two band 2 schools.

Overall, the finding that students from two schools of similar banding level demonstrated similar performance on the two assessment task types provides empirical support for reliability and fairness of the SBA as a component in the public assessment of the English language subject at the secondary level in Hong Kong. Meanwhile, the possibility that teacher rater's leniency might lead to higher scores in some domains of group discussion task in the band 3 school in this study provides justification to the need for the HKEAA to adopt a statistical moderation method in moderating the SBA assessment results submitted by schools to ensure the comparability of SBA scores across schools. Finally, observation of an obviously more fine-grained pattern of difference in difficulty levels of different assessment domains in students' assessment results across the two task types clearly adds to our understanding of the role of different task types in oral assessment in the classroom assessment context. The generalizability of the specific results of this study, however, could be limited by its small sample of schools involved in this study. Future studies should use a more representative sample of schools selected from a variety of geographic regions across the region.

Appendix 1

Table 2 SBA assessment criteria for group interaction (GI)

	I. Pronunciation and delivery	II. Communication strategies	III. Vocabulary and language patterns	IV. Ideas and organization
6	Can project the voice appropriately for the context without artificial aids. Can pronounce all sounds/sound clusters and words clearly and accurately. Can speak fluently and naturally, with very little hesitation, while using suitable intonation to enhance communication.	Can use appropriate body language to display and encourage interest. Can use a full range of turn-taking strategies to initiate and maintain ap propriate interaction, and can draw others into the interaction (e.g., by summarizing for weaker students' benefit or by redirecting a conversation to a quiet student). Can interact without the use of narrowly formulaic expressions.	Can use a wide range of accurate and appropriate vocabulary. Can use varied, appropriate, and highly accurate language patterns; minor slips do not impede communication. Can self-correct effectively. May occasionally glance at notes but is clearly not dependent on them.	Can express a wide range of relevant information and ideas without any signs of difficulty and without the use of notes. Can consistently respond effectively to others, sustaining and extending a conversational exchange. Can use the full range of questioning and response levels (see Framework of Guiding Questions) to engage with peers.
5	Can project the voice appropriately for the context without artificial aids. Can pronounce all sounds/sound clusters clearly and almost all words accurately. Can speak fluently using intonation to enhance communication, with only occasional hesitation, giving an overall sense of natural non-native language.	Can use appropriate body language to display and encourage interest. Can use a good range of turn-taking strategies to initiate and maintain appropriate interaction and can help draw others into the interaction (e.g., by encouraging contributions, asking for opinions, or by responding to group members' questions). Can mostly interact without the use of narrowly formulaic expressions.	Can use varied and almost always appropriate vocabulary. Can use almost entirely accurate and appropriate language patterns. Can usually self-correct effectively. May occasionally refer to a note card.	Can express relevant information and ideas clearly and fluently, perhaps with occasional, unobtrusive, reference to a notecard. Can respond appropriately to others to sustain and extend a conversational exchange. Can use a good variety of questioning and response levels (see Framework of Guiding Questions).
4	Can project the voice mostly satisfactorily without artificial aids. Can pronounce most sounds/sound clusters and all common words clearly and accurately; less common words can be understood although there may be articulation errors (e.g., dropping final consonants). Can speak at a deliberate pace, with some hesitation but using sufficient intonation conventions to convey meaning.	Can use some features of appropriate body language encourage to and display interest. Can use a range of appropriate turn-taking strategies to participate in interaction (e.g., by making suggestions in a group discussion), and can sometimes help draw others in (e.g., by asking for their views). Can interact using a mixture of mainly natural language and formulaic expressions.	Can use mostly appropriate vocabulary. Can use language patterns that are usually accurate and without errors that impede communication. Can self-correct when concentrating carefully or when asked to do so. May refer to a note card but is not dependent on notes.	Can present relevant literal ideas clearly in a well-organized structure, perhaps with occasional reference to a notecard. Can often respond appropriately to others; can sustain and may extend some conversational exchanges However, can do these things less well when attempting to respond to interpretive or critical questions, or when trying to interpret information and present elaborated ideas.
3	Volume may be a problem without artificial aids. Can pronounce all simple sounds clearly but some errors with sound clusters; less common words may be misunderstood unless supported by contextual meaning.	Can use appropriate body language to display interest in the interaction. Can use appropriate but simple turn-taking strategies to participate in, and occasionally initiate, interaction (e.g., by requesting repetition and clarification or by offering agreement).	Can use simple vocabulary and language patterns appropriately and with errors that only occasionally impede communication. Can sometimes self-correct simple errors. May suggest a level	Can present some relevant ideas sequentially with some links among own ideas and with those presented by others. Can respond to some simple questions and may be able to

Table 2 SBA assessment criteria for group interaction (GI) *(Continued)*

	I. Pronunciation and delivery	II. Communication strategies	III. Vocabulary and language patterns	IV. Ideas and organization
	Can speak at a careful pace and use sufficient basic intonation conventions to be understood by a familiar and supportive listener; hesitation is present.	Can use mainly formulaic expressions as communication strategies.	of proficiency above 3 but has provided too limited a sample, or cannot be scored accurately because of dependence on notes.	expand these responses when addressed directly.
2	Volume may be a problem without artificial aids. Can pronounce simple sounds/sound clusters well enough to be understood most of the time; common words can usually be understood within overall context. Can produce familiar stretches of language with sufficiently appropriate pacing and intonation to help listener's understanding.	Can use appropriate body language when especially interested in the group discussion or when prompted to respond by a group member. Can use simple but heavily formulaic expressions to respond to others (e.g., by offering greetings or apologies).	Can appropriately use vocabulary drawn from a limited and very familiar range. Can use some very basic language patterns accurately in brief exchanges. Can identify some errors but may be unable to self-correct. Provides a limited language sample or a sample wholly spoken from notes.	Can express some simple relevant information and ideas, sometimes successfully, and may expand some responses briefly. Can make some contribution to a conversation when prompted.
1	Volume is likely to be a problem. Can pronounce some simple sounds and common words accurately enough to be understood. Can use appropriate intonation in the most familiar of words and phrases; hesitant speech makes the listener's task difficult.	Can use restricted features of body language when required to respond to peers. Can use only simple and narrowly restricted formulaic expressions and only to respond to others.	Can produce a narrow range of simple vocabulary. Can use a narrow range of language patterns in very short and rehearsed utterances. The language sample is too limited for a full assessment of proficiency.	Can occasionally produce brief information and ideas relevant to the topic. Can make some brief responses or statements made when prompted.
0	Does not produce any comprehensible English speech.	Does not use any interactional strategies.	Does not produce any recognizable words or sequences.	Does not produce any appropriate, relevant material.

Appendix 2

Table 3 SBA assessment criteria for individual presentation (IP)

	I. Pronunciation and delivery	II. Communication strategies	III. Vocabulary and language patterns	IV. Ideas and organization
6	Can project the voice appropriately for the context without artificial aids. Can pronounce all sounds/sound clusters and words clearly and accurately. Can speak fluently and naturally, with very little hesitation, while using suitable intonation to enhance communication.	Can use appropriate body language to show focus on audience and to engage interest. Can judge timing in order to complete the presentation. Can confidently invite and respond to questions if this is required by the task.	Can use a wide range of accurate and appropriate vocabulary. Can use varied, appropriate and highly accurate language patterns; minor slips do not impede communication. Can choose appropriate content and level of language to enable audience to follow. Can self-correct effectively. Can present without use of notes, but may glance at a note card occasionally.	Can convey relevant information and ideas clearly and fluently without referring to notes. Can elaborate in detail on some appropriate aspects of the topic, and can consistently link main points with support and development. Can be followed easily and with interest. Can reformulate a point if the audience is unclear.
5	Can project the voice appropriately for the context without artificial aids. Can pronounce all sounds/sound clusters clearly and almost all words accurately. Can speak fluently using intonation to enhance communication, with only occasional hesitation, giving an overall sense of natural nonnative language.	Can use appropriate body language to show focus on audience and to engage interest. Can judge timing sufficiently to cover all essential points of the topic. Can appropriately invite and respond to questions or comments when required for the task.	Can use varied and almost always appropriate vocabulary. Can use almost entirely accurate and appropriate language patterns. Can choose content and level of language that the audience can follow, with little or no dependence on notes. Can usually self-correct effectively. May occasionally refer to a note card.	Can convey relevant information and ideas clearly and well, perhaps with occasional, unobtrusive, reference to a note card. Can elaborate on some appropriate aspects of the topic, and can link main points with support and development. Can be followed easily. Can explain a point if the audience is unclear.
4	Can project the voice mostly satisfactorily without artificial aids. Can pronounce most sounds/sound clusters and all common words clearly and accurately; less common words can be understood although there may be articulation errors (e.g., dropping final consonants). Can speak at a deliberate pace, with some hesitation but using sufficient intonation conventions to convey meaning.	Can use appropriate body language to display audience awareness and to engage interest, but this is not consistently demonstrated. Can use the available time to adequately cover all the most essential points of the topic. Can respond to any well-formulated questions if these are required by and directly related to the task	Can use mostly appropriate vocabulary. Can use language patterns that are usually accurate and without errors that impede communication. Can choose mostly appropriate content and level of language to enable audience to follow. Can self-correct when concentrating carefully or when asked to do so. May refer to a note card but is not dependent on notes.	Can present relevant literal ideas clearly in a well-organized structure, perhaps with occasional reference to a note card. Can expand on some appropriate aspects of the topic with additional detail or explanation, and can sometimes link these main points and expansions together effectively. Can be followed without much effort.
3	Volume may be a problem without artificial aids. Can pronounce all simple sounds clearly but some errors with sound clusters; less common words may be misunderstood unless supported by contextual meaning. Can speak at a careful pace and use sufficient basic intonation conventions to be understood	Can use some appropriate body language, displaying occasional audience awareness and providing some degree of interest. Can present basic relevant points but has difficulty sustaining a presentation mode. Can respond to any relevant, cognitively simple, well-formulated questions required by the task.	Can use simple vocabulary and language patterns appropriately and with errors that only occasionally impede communication, but reliance on memorized materials or written notes makes language and vocabulary use seem more like written text spoken aloud. Can choose a level of content and language that enables audience to follow a main point, but	Can present some relevant literal ideas clearly, and can sometimes provide some simple supporting ideas. Can sometimes link main and supporting points together. May appear dependent on notes.

Table 3 SBA assessment criteria for individual presentation (IP) *(Continued)*

	I. Pronunciation and delivery	II. Communication strategies	III. Vocabulary and language patterns	IV. Ideas and organization
	by a familiar and supportive listener; hesitation is present.		needs to refer to notes. Can sometimes self-correct simple errors, may suggest a level of proficiency above 3, but cannot be scored accurately because of dependence on notes.	
2	Volume may be a problem without artificial aids. Can pronounce simple sounds/sound clusters well enough to be understood most of the time; common words can usually be understood within overall context. Can produce familiar stretches of language with sufficiently appropriate pacing and intonation to help listener's understanding.	Can use a restricted range of features of body language, but the overall impression is stilted. Can present very basic points but does not demonstrate use of a presentation mode and is dependent on notes. Audience awareness is very limited.	Can appropriately use vocabulary and language patterns drawn from a limited and very familiar range. Can read notes aloud but with difficulty. Can identify some errors but may be unable to self-correct. Provides a limited language sample or a sample wholly spoken from notes.	Can make an attempt to express simple relevant information and ideas, sometimes successfully, and can attempt to expand on one or two points. Can link the key information sequentially. May be dependent on notes.
1	Volume is likely to be a problem. Can pronounce some simple sounds and common words accurately enough to be understood. Can use appropriate intonation in the most familiar of words and phrases; hesitant speech makes the listener's task difficult.	Body language may be intermittently present, but communication strategies appropriate to delivering a presentation are absent. There is no evident audience awareness.	Can produce a narrow range of simple vocabulary. Can use a narrow range of language patterns in very short and rehearsed utterances. Insufficient sample to assess vocabulary and language patterns.	Can express a main point or make a brief statement when prompted, in a way that is partially understandable. The presentation is wholly dependent on notes or a written text.
0	Does not produce any comprehensible English speech.	Does not attempt a presentation.	Does not produce any recognizable words or sequences.	Does not express any relevant or understandable information.

Acknowledgements
We would like to thank Professor Antony John Kunnan for his helpful comments on an earlier version of the paper.

Authors' contributions
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 Education, University of Macau, Macao, People's Republic of China. [2]Faculty of Arts and Social Sciences, UNSW, Sydney, NSW 2052, Australia.

Received: 24 July 2017 Accepted: 2 November 2017
Published online: 04 December 2017

References

American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
Bachman, L, Lynch, B, Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238–257.
Bachman, L.F. and Palmer, A.S. 1996: Language testing in practice: designing and developing useful language tests. Oxford: Oxford University Press.
Bond, TG, & Fox, CM (2015). *Applying the Rasch model*, (3rd ed.,). New York and London: Routledge, Taylor & Francis Group.
Boone, W.J., Staver, J.R., & Yale, M.S. (2014). Rasch Analysis in the Human Sciences. Dordrecht: Springer.
Bonk, WJ, & Ockey, GJ. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.
Bygate, M. (1999). Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research*, *3*(2), 185–214.
Carless, DR, & Harfitt, G (2013). Innovation in secondary education: A case of curriculum reform in Hong Kong. In K Hyland, LC Wong (Eds.), *Innovation and change in English language education*, (pp. 172–185). London: Routledge.
Chiedu, RE, & Omenogor, HD. (2014). The concept of reliability in language testing: issues and solutions. *Journal of Resourcefulness and Distinction*, *8*(1), 1–9.
Choi, CC. (1999). Public examinations in Hong Kong. *Assessment in Education*, *6*(3), 405–418.
Clapham, C. (2000) Assessment and testing. *Annual Review of Applied Linguistics*, 20:147–161.
Crooks, TJ. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438–481.
Davison, C. (2007). Views from the chalk face: school-based assessment in Hong Kong. *Language Assessment Quarterly*, *4*(1), 37–68.
Davison, C., Hamp-Lyons, L., Leung, W., Gan, Z., Poon C., Fung, V. 2010. Longitudinal Study on the Schoolsbased Assessment Component of the 2007 Hong Kong Certificate of Education (HKCE) English Language Examination. The Hong Kong Examinations and Assessment Authority.
Ellis, R (1994). *The study of second language acquisition*. Oxford: Oxford University Press.
Foster, P, & Skehan, P. (1996). The influence of planning time and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299–323.
Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing*, *13*, 23–51.
Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Language Assessment Quarterly*, *9*(2), 133–151.
Gan, Z. (2013). Task type and linguistic performance in school-based assessment situation. *Linguistics and Education*, *24*, 535–544.
Harlen, W. (2005). Trusting teachers' judgement: research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, *20*(3), 245–270.
Hong Kong Examination and Assessment Authority (HKEAA). (2016). English language school-based assessment teachers' handbook.
Kane, M. (2010). Validity and fairness. *Language Testing*, *27*, 177–182.
Linacre, J.M. (2014). WINTSEPS (Version 3.81.0) [Computer Software]. Retrieved 17 November, 2016 from Chicago: winsteps.com.
Linacre, J. M. (2017). Facets computer program for many-facet Rasch measurement, version 3.80.0. Retrieved 20, October 2017 from Beaverton, Oregon: Winsteps.com.
Liu, X (2010). *Using and developing measurement instruments in science education: a Rasch modeling approach*. Charlotte: Information Age Publishing.
McNamara, T (1996). *Measuring second language performance*. London & New York: Longman.
Meisels, SJ, Bickel, DD, Nicholson, J, Xue, Y, Atkins-Burnett, S. (2001). Trusting teachers' judgments: a validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal*, *38*(1), 73–95.
Michel, MC, Kuiken, F, Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, *45*(2), 241–259.

Price, M, O'Donovan, B, Rust, C. (2007). Putting a social-constructivist assessment process model into practice: building the feedback loop into the assessment process through peer review. *Innovations in Education and Teaching International*, *44*(2), 143–152.
Qian, DD. (2014). School-based English language assessment as a high-stakes examination component in Hong Kong: insights of frontline assessors. *Assessment in Education: Principles, Policy & Practice*, *21*(3), 251–270.
Rasch, G (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: The University of Chicago Press.
Rahimpour, M., 2007. Task complexity and variation in L2 learners' oral discourse.The University of Queensland Working papers in Linguistics, Australia., 1: 1–9.
Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: effects on L2 speech production, interaction, uptake, and perceptions of task difficulty. *International Review of Applied Linguistics*, *45*(3), 193–213.
Robinson, P, Cadierno, T, Shirai, Y. (2009). Time and motion: measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics*, *30*(4), 533–544.
Sadler, D.R. (2006). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, 5 (1):77–84.
Skehan, P. (2009). Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510–532.
Skehan, P (Ed.) (2014). *Processing perspectives on task performance*. London: John Benjamins.
Skehan, P, & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research*, *1*(3), 185–211.
Tarone, E. (1990). On variation in interlanguage: a response to Gregg. *Applied Linguistics*, *11*, 392–400.
Vygotsky, L (1978). *Mind in society: the development of higher psychological processes*. Cambridge: Harvard University Press.
Wright, BD (1999). Fundamental measurement for psychology. In SE Embretson, SL Hershberger (Eds.), *The new rules of measurement: what every educator and psychologist should know*, (pp. 65–104). Hillsdale: Lawrence Erlbaum.
Wright, BD, & Linacre, JM. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.
Wright, BD, & Master, GN (1982). *Rating scale analysis*. Chicago: Mesa Press.
Wyatt-Smith, C, Klenowski, V, Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 59–75.
Yip, DY, & Cheung, D. (2005). Teachers' concerns on school-based assessment of practical work. *Journal of Biological Education*, *39*(4), 156–162.