JOURNAL OF APPLIED MEASUREMENT, *17*(3), 335-353

The Self-assessment Practices of Hong Kong Secondary Students: Findings with a New Instrument

Zi Yan
The Hong Kong Institute of Education

Self-assessment is a core skill that enables students to engage in self-regulated learning. The purpose of this study was to examine the psychometric properties of a Self-assessment Practice Scale and to depict the characteristics of self-assessment practices of Hong Kong secondary students using this newly developed instrument. A total of 6,125 students from 10 Hong Kong secondary schools completed the survey. Both Rasch and factor analyses revealed a two-dimension scale structure (i.e., Self-directed Feedback Seeking and Self-reflection). The two subscales demonstrated acceptable psychometric properties and suggestions for further improvement were proposed. The findings regarding self-assessment practices of secondary students indicated that, in general, students were quite used to engaging in self-reflection based on available feedback, but they were less disposed to taking the initiative to seek feedback on their own performance. Key demographic variables, e.g., gender and year level, played important roles in students' self-assessment practices. Girls had significantly higher self-assessment measures on both scales than did boys. Junior students had higher measures on both scales than did their senior counterparts. Implications and directions for future research were discussed.

Requests for reprints should be sent to Zi Yan, Department of Curriculum and Instruction, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, Hong Kong, e-mail: zyan@ied.edu.hk.

Introduction

The current international education reform moving towards a constructivist paradigm requires students to be self-regulated learners who build their own knowledge. Consequentially, an overall aim of the school curriculum in Hong Kong is to help students to learn how to learn and equip students with the ability and disposition to pursue life-long learning (Curriculum Development Council, 2001). One core skill that enables students' engagement in self-regulated learning is self-assessment, by which students take responsibility for reflecting on and evaluating their own learning progress and/or products (Kirby and Downs, 2007; Langendyk, 2006). Compared to other assessments, in particular tests and examinations, which often have negative consequences for students' leaning attitudes and behaviours, self-assessment has the potential to improve students' commitment to and autonomy in learning (Paris and Paris, 2001).

Perhaps largely due to its name, self-assessment, the majority of research on the topic has focused on its assessment function. Unfortunately, many studies have revealed that self-assessment is inaccurate, and an unreliable indicator of student performance (Brown and Harris, 2013; Ross, 2006); and that inaccuracy of self-assessment is somewhat a by-product of a human disposition of being unrealistic about one's own abilities (Dunning, Heath, and Suls, 2004). On the other hand, Boud (1999) argued that self-assessment can contribute more to learning by acting as a learning process rather than as a substitute for other-conducted assessments. In other words, engagement in self-assessment itself, rather than providing accurate self evaluations, is more important because engaging in self-assessment results in learning gains by enhancing students' self-regulation and metacognitive skills (Kostons van Gog, and Paas, 2012; Panadero, Alonso-Tapia, and Huertas, 2012). Therefore, it is self-assessment practice, rather than the scores or ratings generated by self-assessment, that really matters in terms of improving students' performance. A realistic approach in developing students to be self-regulated learners is to enhance teachable self-assessment practices, such as feedback seeking and self-reflection, rather than requiring accurate self-judgment.

Despite the important role self-assessment practices can play in students' learning, a detailed description of its characteristics is missing in literature. For example, which kind(s) of activities do students undertake in their self-assessment? What is the prevalence of self-assessment practices? Are there any differences in self-assessment practices for different groups of students (e.g., by gender, or year level)? This study aimed to fill this gap.

The conceptualizations and roles of self-assessment

In the self-assessment literature, it seems that there are quite different conceptualizations and interpretations of what actually constitutes self-assessment. These conceptualizations could be summarized into three major groups. First is that self-assessment is conceptualized as the personal ability/skill of carefully considering the quality of one's own work and making a summative evaluation of one's own knowledge, skills, or performance (e.g., Eva and Regehr, 2008; Noonan and Duncan, 2005). The second group views self-assessment as a type of assessment satisfying summative purposes (e.g., Boud, 1999; Bosma, Laszakovits, and Hattery, 2007). The third group conceptualizes self-assessment as a learning strategy or process which promotes students' learning (e.g., Panadero and Alonso-Tapia, 2013; Paris and Paris, 2001; Puustinen and Pulkkinen, 2001; Zimmerman and Moylan, 2009).

The implications of viewing self-assessment as a personal ability/skill are limited as Eva and Regehr (2008) asserted that self-assessment is unlikely to be developed as a personal skill; inaccuracy of self-assessment being a natural consequence of unrealistic human nature (Dunning, Heath, and Suls, 2004). This position has been supported in a large number of studies (e.g., Harris and Brown, 2010; Kuncel, Credé, and Thomas, 2005; Powel and Gray, 1995).

From the perspective of assessment, Ross (2006) concluded, in a review study, that self-

assessment was a less reliable indicator of student performance than were other assessments (e.g., test scores, teacher/parent ratings). A more recent review of self-assessment (Brown and Harris, 2013) echoed that, in general, self-assessment was not robustly accurate and that caution is warranted in using self-assessment results for accountability purposes due to its inaccuracy, especially for younger or less able students. Blackwood (2010) further reported that overconfidence in self-assessment is particularly salient for Chinese students. The inaccuracy of self-assessment, along with its detachment from reality, attenuated the status of self-assessment as a useful type of assessment (Brown and Harris, 2014) since assessment practices must be valid and reliable in order to inform decision-making (Messick, 1989).

Although self-assessment, as a personal skill or a type of summative assessment, seems inherently flawed and the associated assessment value is therefore diminished, that should not dismiss the possible merits of self-assessment as a key learning strategy. From a pedagogical perspective, self-assessment encourages students to reflect on their own performances and enhances their self-regulation which, in turn, leads to improved learning. In this sense, the importance of the accuracy of self-assessment seems reduced as Panadero, Brown, and Strijbos (2014, p. 11) argued that "While it seems appropriate that an assessment be accurate, it might well be that there are benefits from engaging in self-assessment, even if it is unrealistic or inaccurate as long as students overcome any unrealism or inaccuracy." Following this line of reasoning, the self-assessment practices students engage in, rather than making a final evaluation, appear crucial for learning improvement.

The components of self-assessment practices

When the need for self-assessment arises, the first thing students have to do is to gather information that gives them evidence as to the quality of their performances. According to the availability of such information/feedback, the current study differentiated students' self-assessment practices into two classes, namely, self-directed feedback seeking and self-reflection. When the feedback regarding students' performances—such as teachers' written comments on the assignments, marks on exam papers, etc.—are available, the self-assessment practices students conduct are related to self-reflection. Reflection refers to mental endeavor aiming at exploring and elaborating one's understanding of problems encounter during learning (Mann, Gordon, and McLeod, 2009). Reflective thinking is crucial component of self-assessment process (McMillan and Hearn, 2008).This activity helps students to understand better the learning process and to identify their strengths and weaknesses. By reflection students can obtain information on what they learned, what they need to work on, and how they can achieve the goal.

However, if the information regarding students' performances are missing or the meaning of such information is blurred, students need to take responsibility for seeking feedback about the quality of their performances (Boud, 1999; Kirby and Downs, 2007; Langendyk, 2006). Although the most accessible feedback source is from students themselves, e.g., self-testing, self-rating one's own performance, or self-recording learning progress, Boud (1999, p. 122) argued that "the practice of self-assessment does not imply that this engagement is an isolated or individualistic activity. It commonly involves peers, makes use of teachers and other practitioners and draws upon appropriate literature." In his conceptualization, self-assessment practice involves a process by which one takes initiative for seeking feedback from various sources including teachers, peers, and parents, though the role and the degree of involvement of external parties may vary. Thus, self-directed feedback seeking is another indispensable component of self-assessment practices.

In this study, the focus is on self-assessment practices of Hong Kong secondary students. There are two purposes in this study. One is to develop a scale to assess student self-assessment practices and to provide a preliminary examination of the psychometric properties of that scale. The second is to detail the characteristics of

Hong Kong secondary students' self-assessment practices. The findings should help researchers and teachers to better understand and promote students' self-assessment practices which lead to improved learning outcomes.

Methodology

The methodology section of this paper consists of two parts: the procedure for scale development and the subsequent administration of the scale. The research sample and data analysis methods used are described in the latter section.

Scale development

Although self-assessment could be understood in a subject-based context, the self-assessment practice in this study refers to those activities in which students tend to engage across different subject areas. The initial item pool was built from descriptions about self-assessment practices which came from two major sources: (1) available instruments relevant to self-assessment or self-regulated learning (e.g., Motivated Strategies for Learning Questionnaire developed by Pintrich, Smith, Garcia, and Mckeachie, 1991; the Self-directed Learning Scale developed by Mok, Cheng, Moore, and Kennedy, 2006); as well as (2) consultative discussions with teachers ($N = 6$) and secondary students ($N = 25$). The identified self-assessment practices were formulated as 18 statements. These statements were then subjected to an intensive review by a panel of three experts in the field of educational assessment to examine the face and content validity of the statements. Experts rated the appropriateness of each statement in a scale (yes/no/unsure) independently. The statement was retained only if the majority of panel (2 or 3 experts) agreed that the item was appropriate to be used to gauge self-assessment practice. This process resulted in a 16-item scale. The expert panel was also asked to review the clarity, possible ambiguity, and potential bias of each of the remaining items. Necessary modifications were made. The revised 16 items were then presented to another group of secondary students ($N = 25$) who were asked to select their most frequently-used self-assessment practices. The frequencies were tallied and the 10 most frequently endorsed items were included in the final scale (see Appendix 1). The 10-item scale, instead of the 16-item scale, was used because of two reasons. First, this scale had to maintain a minimum number of items since it was part of a survey administered in a large project. Second, the 6 items being removed from the final scale received significantly less endorsement from the pilot group of students than the 10 items being kept. The frequencies for the 10 items in the final scale ranged from 48% to 88%, while the frequencies for the 6 items being removed ranged only from 8% to 12%. Among the 10 items, 7 items are about self-directed feedback seeking and 3 items are related to self-reflection. Respondents to the scale were asked to indicate their level of agreement with each statement on a six-point Likert-type scale, ranging from *Strongly Disagree* (1), *Disagree* (2), *Slightly Disagree* (3), *Slightly Agree* (4), *Agree* (5) to *Strongly Agree* (6). Given that the questionnaires were to be completed by Hong Kong local students, items were developed in Cantonese.

Sample

The survey was administered in 10 Hong Kong secondary schools that had participated in a Quality Education Fund project supported by the Hong Kong government. The10 participating schools were carefully selected with a stratified sampling method using school bands as the stratum to ensure that the participating students represented a wide range of academic ability. All students in the participating schools were invited to participate in the survey. Among the 6,125 respondents who completed the questionnaire, there were 2,971 (48.5%) males, 3,074 (50.2%) females, and 80 (1.3%) who omitted that information. The sample consists of students across all secondary school year levels from S1 (983, 16.0%), S2 (1092, 17.8%), S3 (1045, 17.1%), S4 (1089, 17.8%), S5 (1206, 19.7%), to S6 (710, 11.6%).

Data analysis

In fulfilling the first research purpose, both Rasch analysis (Rasch, 1960) and factor analy-

sis were used to investigate the psychometric properties of the scale. For the second research purpose, descriptive analyses and *t*-test/ANOVA were applied to students' Rasch-calibrated self-assessment measures to reveal (a) the range and the distribution of the students' responses to those items, and (b) the differences in self-assessment practices by gender and year level.

With the principle that an objective measurement should target only one attribute or dimension at one time (Bond and Fox, 2015) as a tenet, Rasch analysis is used to check the extent to which items in any scale belong to a single unidimensional construct. Once the unidimensional scale is built, person ability and item difficulty can be calibrated onto a single latent trait scale. Persons/items are placed on an ordered trait continuum according to their ability/difficulty estimated by the Rasch model. Therefore the meaning of person measures and item difficulties can be interpreted under the same framework and direct comparisons between person measures and item difficulties can be conducted. According to the Rasch model, the outcome of a person's answer to a given item is determined by the comparison of person ability and item difficulty. The higher a person's ability compared to the difficulty of the item, the higher the probability that s/he will have a right answer to the item. Mathematically, the person ability measure and item difficulty measure are estimated in the same way in the Rasch model. First, the ratio of each person's percentage of correct answers over the percentage of incorrect answers is calculated, and then that ratio is transformed into odds of a successful response. Finally, the natural logarithm of those odds is calculated as the person ability measure (or, similarly, the item difficulty measure). Through several rounds of iteration, the raw data are transformed into interval measures (in logits) of person ability and item difficulty (Bond and Fox, 2015).

With the Rasch model, a "data fit the model" approach is employed (Andrich, 2004). This approach requires that the empirical data must meet specific requirements for the purpose of fundamental measurement. In current study, the criteria used to examine the data-model fit include response category functioning, dimensionality investigations, Rasch separability reliability, and item fit statistics. The standardized form of the outfit mean square, i.e., Outfit ZSTD, was used as the item fit statistic in this study. In many studies, items were regarded as misfitting to the Rasch model if the Outfit ZSTD was larger than +2.0. However, a relatively lenient critical value, i.e., +3.0, could be a more meaningful criterion for large samples (Linacre, 2002). Given that the sample used in this study was over 6,000, +3.0 was adopted as the critical value of Outfit ZSTD for detecting misfit items. Differential Item Functioning (DIF) across gender and year level was also checked since such key demographic variables were found to have substantial impact on self-regulated learning activities or readiness in previous studies (e.g., Bidjerano, 2005; Pokay and Blumenfeld, 1990; Reio, 2004; Reio and Davis, 2005). DIF analysis aims to check construct equivalences across groups. DIF exists when an item has different levels of difficulty for certain groups of subjects after controlling their levels of latent trait. As suggested by previous studies (e.g., Wang, Yao, Tsai, Wang, and Hsieh, 2006), a difference equal to or larger than 0.5 logits was considered as substantial DIF in this study.

Factor analysis, on the other hand, usually aims to identify the underlying factor structure and determine which items belong to each factor. Both exploratory factor analysis and confirmatory factor analysis were used in this study. Exploratory factor analysis aims to identify the number and the content of constructs underlying a set of items; while confirmatory factor analysis intends to test the fit between the empirical data and a pre-specified factor structure (Thompson, 2004, p. 6). In this exploratory factor analysis, a factor loading higher than .4 was used as the criterion to select items (Nunnally, 1978). In confirmatory factor analysis, multiple fit indices including chi-square (χ^2), Goodness of Fit Index (GFI), Adjusted Goodness of Fit Index (AGFI), comparative fit index (CFI), the Tucker-Lewis-Index (TLI), and the standardized root-mean-square error of approximation (RMSEA) were used to examine the model data fit.

Although these two analytical methods are built on different assumptions, using both together brings advantage in refining instruments as well as providing a holistic interpretation of research results (Richardson, 2005). For example, confirmatory factor analysis provides numerous overall model fit indices, while Rasch analysis emphasizes more on the fit of each item to the model. Some empirical studies, including a recent investigation of learners in Hong Kong (Deneen, Brown, Bond, and Shroff, 2013) obtained slightly different but complementary results from these two analytical methods. This study attempted to examine whether these two analytical methods, based on the same set of data, would generated different or similar results.

Results

Rasch analysis

Since all items in the scale shared the same six-point rating scale, rating scale Rasch model using WINSTEPS 3.73 (Linacre, 2011) was applied to conduct the Rasch analysis. The category function of the six-point rating scale was first checked. If the categories function well, the threshold calibrations (the intersection points of adjacent category curves) of the rating scale should increase monotonically (Linacre, 2006). It can be seen from the response category probability curves (Figure 1) that the threshold calibrations for the rating scale satisfied this requirement. In other words, a higher measure on the items represents a higher level of latent trait under measurement. Each category had a distinct peak in the graph, indicating each category was the most probable performance level for given groups of persons with a specific level of latent trait.

Putting all the 10 items together, the one-dimension assumption was first checked. One self-reflection item (1: *When the teachers ask the others, I will think about the answers*) showed substantial misfit to the Rasch model. The Outfit ZSTD of this item (+9.9) was higher than +3.0. After removing item 1, another self-reflection item (7: *I will try to figure out the reasons for the mistakes I made after receiving the marked exam paper*) appeared misfitting. The Outfit ZSTD was +8.9. After removing items 1 and 7, the results of Rasch analysis showed that all remaining 8 items in the scale fit the unidimensional Rasch model quite well.

The dimensionality investigations are based on a map generated by WINSTEPS plotting the results of principal components analysis (PCA) of residuals. The results in Figure 2 revealed that

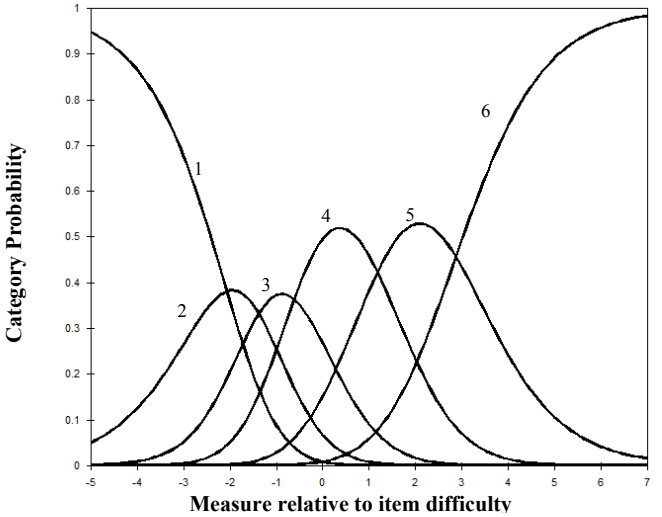

Figure 1. The response category probability curves for the 6-point rating scale.

these two misfitting items (1 and 7), together with item 8 (*I will identify the areas for improvement according to the teacher's comment on my assignment*), which is also a self-reflection item, constituted a second dimension other than the Rasch dimension representing the other self-assessment items. The dimensionality map for this scale showed that the loadings of three items (1, 7, and 8) on the first contrast (.50, .76., and .64 respectively) were substantially different from the loadings of other items (e.g., the loadings of items 2, 3, and 10 are −.34, −.45, and −.33 respectively). It indicated that there were contrasting patterns in the residuals. In other words, while other items represent the main Rasch dimension, these three items come from a separate dimension.

The results of this dimensionality investigation conformed to the 2-dimension conceptual framework of self-assessment practices discussed earlier. Therefore, the 10 items were divided into two scales. One was a *Self-reflection* scale of 3 items (1, 7, and 8), gauging students' self-initiated activities to reflect on their performances based on available feedback. The other was a *Self-directed Feedback Seeking* scale, which was comprised of 7 items (2, 3, 4, 5, 6, 9, and 10), assessing students' self-initiated activities to search for feedback on their performances from a variety of sources.

The Rasch analyses were then undertaken separately for the two scales. The Rasch person reliabilities for these two scales were 0.89 and 0.78 respectively. Over half (57% and 60% respectively) of variance in the data was explained by the Rasch measures in each of these two scales. The items measures, Outfit ZSTDs, DIF across gender and year level are presented in Table 1.

It can be seen from Table 1 that the Outfit ZSTDs for all items in the two scales fall into the acceptable range (lower than +3.0) except that the Outfit ZSTD for item 1 was marginal (3.2). Although there were some DIF effects with statistical significance ($p < .05$) across year level, no DIF value exceeded 0.5 logits which is suggested as a cut-off value of substantive DIF according to Linacre (2006). Given that large sample size (e.g., $N = 6{,}125$ in this study) might flag trivial differences as statistically significant, these DIF results were considered as not substantively meaningful. In sum, all the criteria including Rasch separability reliability, variance explained by measures, item fit statistics, and DIF gave credence to the

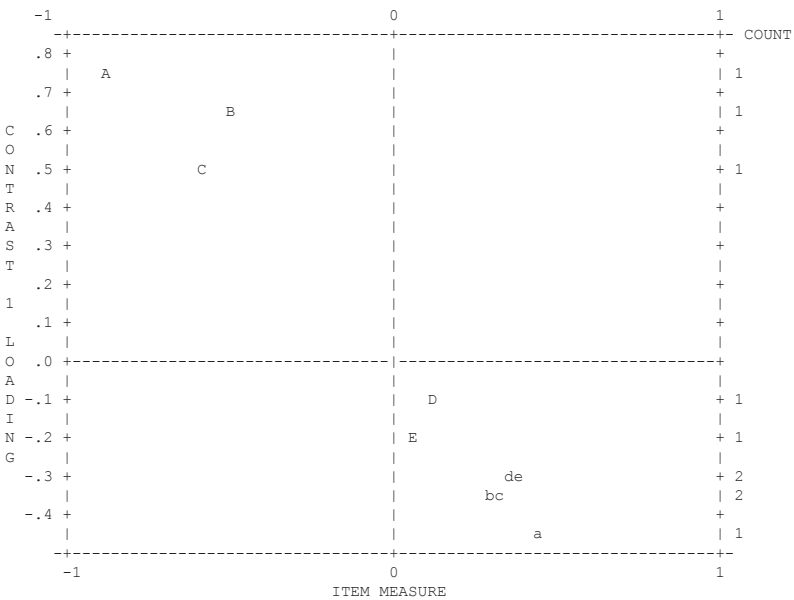

Figure 2. The standardized residual contrast 1 plot.

Table 1

Item Measures, Outfit ZSTDs, DIF across Gender and Year level for the Two Scales.

			DIF	
Scale/Item	Measure	Outfit ZSTD	Gender[1]	Level[2]
Self-directed feedback seeking				
Item #2	0.03	2.9	−0.08	(−0.16, 0.19)
Item #3	0.22	−8.9	0.00	(−0.03, 0.04)
Item #4	−0.27	−5.2	0.12	(−0.24, 0.15)
Item #5	0.10	2.5	0.10	(−0.06, 0.10)
Item #6	−0.19	−6.6	−0.10	(−0.05, 0.07)
Item #9	0.10	2.8	0.00	(−0.23, 0.19)
Item #10	0.02	−0.8	0.00	(−0.09, 0.10)
Self-reflection				
Item #1	0.11	3.2	−0.02	(−0.19, 0.13)
Item #7	−0.30	−6.2	0.00	(−0.03, 0.02)
Item #8	0.19	−9.0	0.06	(−0.08, 0.16)

Note. [1] The figures in this column are the item difficulty differences, in logits, between males and females.

[2] The figures in parentheses in this column are the range of item difficulty differences, in logits, among different year levels. For example, (−0.16, 0.19) for item 1 means that the item difficulty differences between each year level and its average difficulty (baseline) range from −0.16 to 0.19 logits.

appropriateness of using this instrument as two unidimensional scales with the current sample.

Factor analysis

The percentage of missing data for the 10 items ranged from 0.1% to 12.4%. According to Allison's (2003) recommendation, maximum likelihood methods with the expectation maximisation algorithm was used to impute missing data. The dataset was randomly split into two half: one for exploratory factor analysis, the other for confirmatory factor analysis. Exploratory factor analysis, using maximum likelihood estimation with direct oblique rotation, was first conducted on the first half of the sample to explore the factor structure of the data set. Using eigen values > 1 as the criterion, the results of analysis identified two factors. The scree test (Cattell, 1966) also supported that factor structure. A factor loading higher than .4 was used as the criterion to select items (Nunnally, 1978). Items 2, 3, 4, 5, 6, 9, and 10 loaded on factor 1 with factor loadings ranging from .66 to .80. Items 1, 7, and 8 loaded on factor 2 with factor loadings ranging from .41 to .93. No item demonstrated cross-loadings on both factors. The factor structure was consistent with the proposed 2-dimension conceptual framework of self-assessment practices as well as the results of the Rasch analysis.

Confirmatory factor analysis was conducted subsequently on the second half of the sample using AMOS 21.0 with a purpose of further assessing how well the data fit the two-factor structure that was identified from the exploratory factor analysis. The results showed that the GFI (0.930), CFI (0.927), and TLI (0.903) indices were all higher than 0.9 except AGFI (0.887). The chi-square ($\chi^2 = 1082.976$, $df = 34$) was significant at the 0.01 level and RMSEA was 0.100. Inspection of modification indices revealed a strong correlation between the residuals of item 9 and 10. In a revised model, the residuals of item 9 and 10 were allowed to correlate and re-analysis was conducted (see Figure 3). The chi-square statistics decreased to 594.037, although it was still significant. This result is not surprising given that chi-square statistics are also quite sensitive to large sample size. RMSEA was improved to 0.074. All other fit statistics were also improved including GFI (0.961), AGFI (0.935), CFI (0.961) and TLI (0.947). These results showed acceptable fit between the empirical data and the proposed factor structure. The correlation between the residuals

of item 9 and 10 (r = .41) could be explained by the strongly-related item contents. Both item 9 (*I will ask my group members to comment on my work in group activities*) and item 10 (*I will invite others to test how well I have mastered the course contents*) are about students' willingness to invite feedback from others on their own performance. Furthermore, as expected, a strong correlation (r = .72) was found between the two factors, i.e., self-directed feedback seeking and self-reflection.

Students' self-assessment practices

The first purpose of this study was to examine the psychometric properties of the scales. The positive results presented above give credence to use of the scales with the current sample. Therefore, the second purpose of this study—to investigate the characteristics of students' responses—is reported so as to contribute to understanding of Hong Kong secondary students' self-assessment practices. This section provides a descriptive analysis of the range and the distribution of the students' self-assessment abilities; and note any similarities and differences of self-assessment practices across gender and year level.

Figures 3 and 4 present the item-person variable maps for the *Self-directed feedback seeking* scale and *Self-reflection* scale. The Rasch model calibrates person measures and item difficulties along the same interval-level latent trait scale. The left-hand side shows the distribution of students along the latent trait scale; while the item difficulties are located by item number on the right-hand side of the scale. The students with the highest self-assessment levels and the items with highest difficulty levels are located at the top of the map, while the students with the lowest levels and the easiest items are located at the bottom. The item response category thresholds are also presented in the map by the notation of *x.y* (e.g., "2.5" stands for the 5th threshold of the item 2).

For the *Self-directed feedback seeking* scale (see Figure 4), the mean of students' abilities (M = −0.10, SD = 1.65) was virtually identical to the mean difficulty of the items (M = .00, SD = 0.16). Student ability ranged from −5.88 to +6.47 logits; while item difficulty ranged from −0.27 to +0.22 logits. The range of item difficulty was much smaller than the range of student ability. However, the ranges of difficulty levels of the

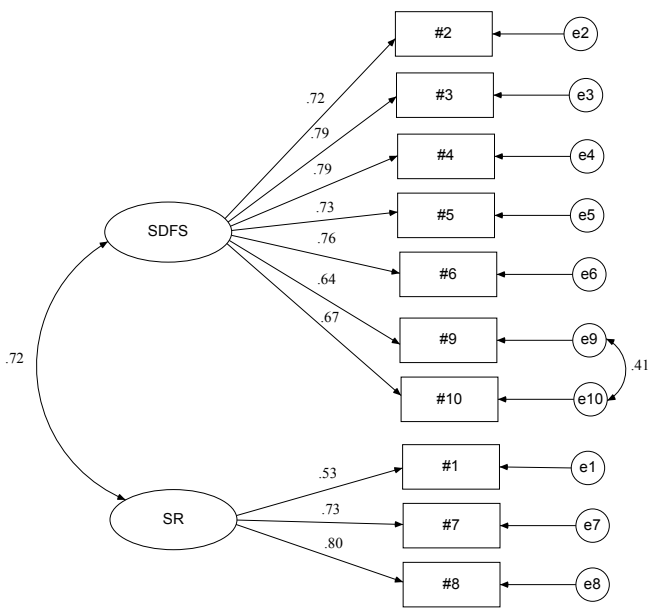

Figure 3. Results of the confirmatory factor analysis on the proposed two-factor model.

response categories for each item, as presented on the right-hand side of the map, covered a wide range of difficulty along the latent trait scale, and provided quite good coverage to the range of student abilities.

For the *Self-reflection* scale (see Figure 5), the mean of students' abilities (M = 0.99, SD = 2.06) was substantially higher than the mean difficulty of the items (M = .00, SD = 0.22). Student ability ranged from −4.73 to +6.56 logits; while item difficulty ranged from −0.30 to +0.19 logits. Similar to the *Self-directed feedback seeking* scale, the items in the *Self-reflection* scale, together with their response categories, covered the range of student abilities.

Comparatively speaking, the items in the *Self-reflection* scale are easier for these secondary school students to endorse: they are more likely to

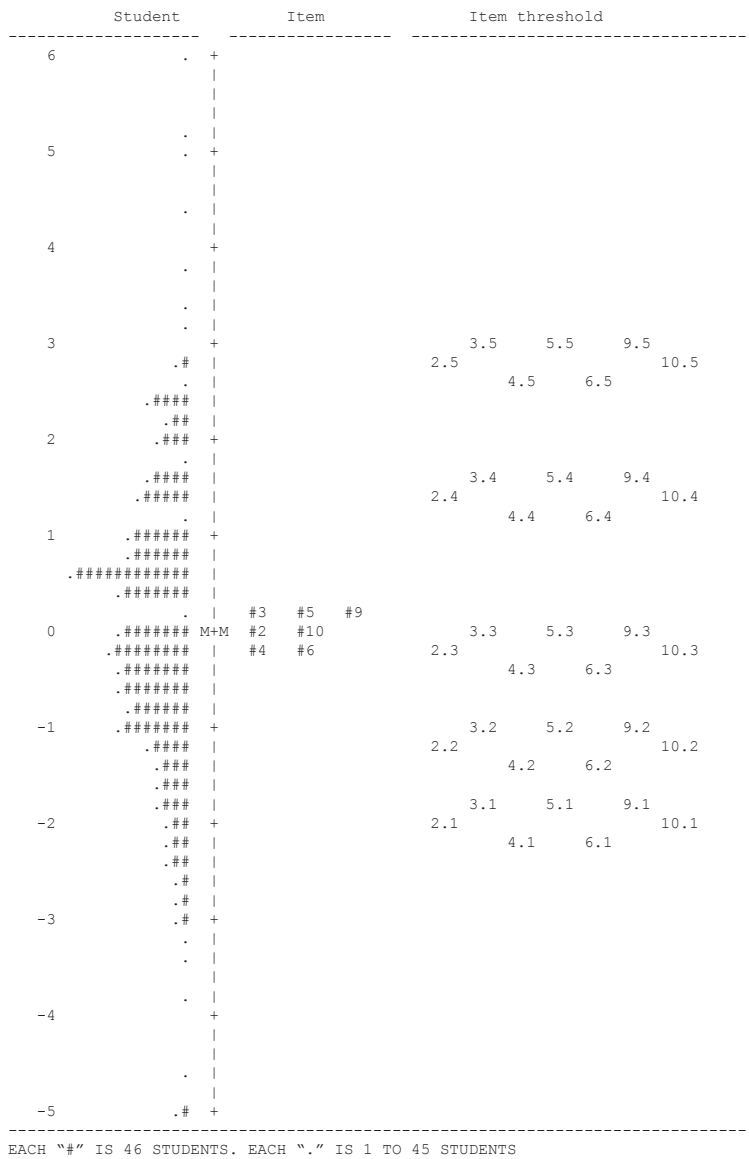

Figure 4. Item-person map for the *Self-directed Feedback Seeking* scale.

engage in self-reflection based on available feedback than to take initiative for seeking feedback from various sources. The most endorsed item was 7 (*I will try to figure out the reasons for the mistakes I made after receiving the marked exam paper*). The most difficult items were on the *Self-directed feedback seeking* scale: 3 (*I often check whether I have fully mastered the course contents through assessing myself using reference books or notes*), 5 (*I will keep track of my progress through the performance record*), and 9 (*I will ask my group members to comment on my work in group activities*).

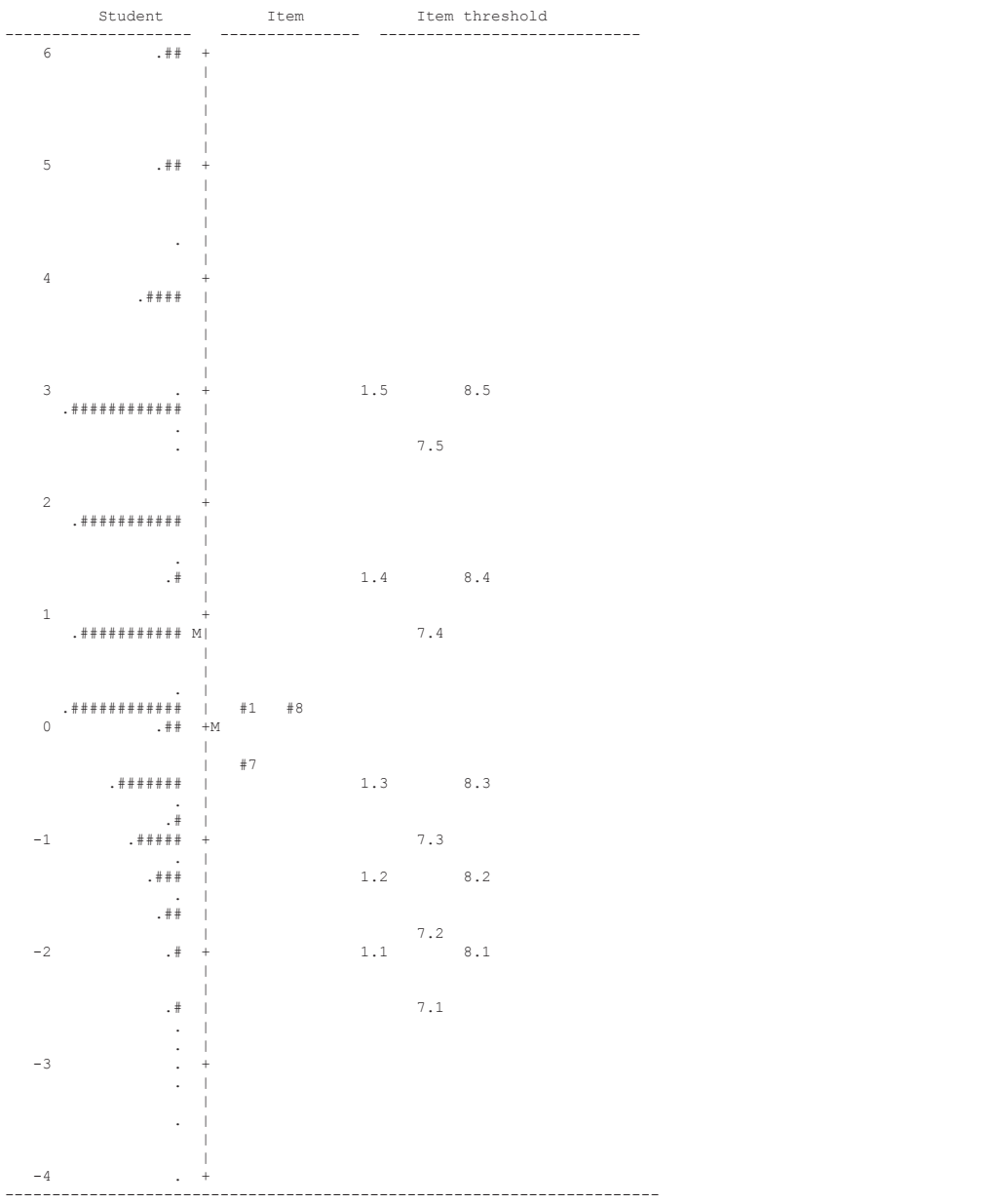

Figure 5. Item-person map for the *Self-reflection* scale.

Table 2 below presents comparisons of Rasch-calibrated mean student measures by gender and year level.

A *t*-test and an ANOVA were undertaken to investigate the differences of self-assessment practices by gender and year level respectively. Table 2 reports a significant ($p < .01$) gender effect on students' responses to both scales. Females, on average, had significantly higher self-directed feedback seeking and self-reflection abilities than did males. The results of ANOVA showed that year level was also an important factor with significant ($p < .01$) effect on students' self-directed feedback seeking and self-reflection abilities: junior students (S1 and S2) reported higher self-assessment abilities than did senior students (S3 to S6) on both scales. The results also showed that there were no significant interaction effects between gender and year level for both scales.

Discussion

The key feature distinguishing self-assessment from other kinds of assessment is that self-assessment is initiated by the student her/himself. Boud (1999) emphasized that self-assessment is a process by which one takes the initiative for seeking feedback on one's own performance. This is in line with the requirement of self-directed learning. According to Knowles (1975, p. 18), self-directed learning is "a process by which individuals take the initiative, with or without the assistance of others, for diagnosing their learning needs, formulating learning goals, identifying human and material resources for learning, choosing and implementing appropriate learning strategies, and evaluating learning outcomes." This study adopted this line of argument and further revealed that such "initiative" in self-assessment practices could be understood on two related, but different psychometric dimensions: those being, seeking feedback and self-reflection.

The first purpose of this study was to examine the psychometric properties of a newly developed self-assessment practice scale for secondary students. Both Rasch and factor analyses resulted in two scales, namely, the *Self-directed Feedback Seeking* scale containing 7 items and the *Self-reflection* scale consisted of 3 items. The resultant two unidimensional scales satisfied the measurement requirements of the Rasch model, i.e., all items in a scale measure a single latent trait, and demonstrated sufficiently good psychometric properties. Both exploratory and confirmatory factor analyses echoed this two-factor structure in students' responses to the items.

This result indicated that students' self-assessment practices should be understood from two perspectives. One dimension concerns students' self-directed activities of seeking feedback on their own performances from different sources. That feedback might come from themselves, including asking themselves questions (2 and 6) and record-keeping (5); from external materials including reference books or notes (3) and past

Table 2

Comparison of Student Measures across Gender and Year Level.

	Self-directed feedback seeking	*Self-reflection*
Gender		
Male ($N = 2971$)	−0.16	0.83
Female ($N = 3074$)	−0.05	1.14
t-test results	$t = -2.593$, $df = 6043$, $p = .010$	$t = -5.771$, $df = 6043$, $p = .000$
Year level		
Secondary 1 (S1) ($N = 983$)	0.02	1.41
Secondary 2 (S2) ($N = 1092$)	−0.04	1.03
Secondary 3 (S3) ($N = 1045$)	−0.15	0.86
Secondary 4 (S4) ($N = 1089$)	−0.03	0.91
Secondary 5 (S5) ($N = 1206$)	−0.26	0.92
Secondary 6 (S6) ($N = 710$)	−0.15	0.77
ANOVA results	$F(5, 6119) = 4.444$, $p = .000$	$F(5, 6119) = 11.446$, $p = .000$

exam papers (4); and from other persons (9 and 10). The other dimension consists of students' reflective activities including self-checking based on teachers' oral feedback (1), figuring out reasons for mistakes on marked exam papers (7) and identifying areas for improvement based on teachers' comments (8). In other words, the major activities students need to undertake to complete self-assessment include gathering feedback as to the quality of their performance from various sources and self-reflection. Thus, self-assessment is not only about evaluating one's own performance according to pre-established criteria, but also about encouraging students to draw on their diverse personal resources to facilitate such evaluation.

The analysis results identified directions of further improvement for the scales. Boud (1999) argued that students' self-assessment practices may involve feedback from various stakeholders. From this perspective, the items included in the *Self-directed Feedback Seeking* scale appear to be not sufficiently comprehensive. For example, parents and teachers are not specifically mentioned in those items. This is partially due to the item selection procedure adopted in this study, i.e. the most frequently used 10 self-assessment practices were included into the final scale. The exclusion of parents or teachers as feedback sources indicated that seeking feedback from parents or teachers might not be popular practice for Hong Kong secondary students. A speculative but possible explanation is that students might not see teachers to be effective in solving their problems (Mok, Kennedy, Moore, Shan, and Leung, 2008). Therefore, students may lack motivation to seek feedback from teachers. However, the inclusion of these feedback sources to provide a more comprehensive picture of students' self-assessment practices still remains open in further development of this scale, as well as in subsequent empirical investigation.

The results also indicated that more items need to be added into the *Self-reflection* scale. The current version of the scale contains only 3 items with lower Rasch person reliability of 0.79. More items would increase the *Self-reflection* scale's reliability and, more importantly, allow more thorough investigation into students' self-reflection activities based on different feedback provided

The satisfactory psychometric properties of the two scales paved the way for the accomplishment of the second purpose of this study, that is, to generate a general picture of the self-assessment practices of Hong Kong secondary students.

Although it is not appropriate to compare directly the difficulty levels of items in these two scales since the items were calibrated in different measurement metrics, the comparison between students' responses did provide information about the relative ease of endorsement of the items in these two scales. The mean of students' abilities was similar to the mean difficulty of the items in the *Self-directed feedback seeking* scale; while substantially higher than the mean difficulty of the items in the *Self-reflection* scale. In other words, students found that the items in the *Self-reflection* scale were relatively easier to endorse, and items in the *Self-directed feedback seeking* scale were relatively difficult to endorse. This finding indicated that students were more used to conducting self-reflection based on available feedback (e.g., teachers' oral feedback and written comments) in order to identify the reasons for mistakes or the areas for improvement.

Previous studies (Panadero, Alonso-Tapia, and Huertas, 2012; Panadero, Alonso-Tapia, and Reche, 2013) revealed that use of self-assessment had positive effect on self-regulation and the magnitude of the effect varied according to different self-assessment tools. In particular, scripts enhanced self-regulation more than rubrics. Panadero, Alonso-Tapia, and Reche (2013) further pointed out that this is because scripts—specific questions designed in steps to analyze the learning process throughout a task—increased metacognitive awareness and, in turn, activated more learning strategies. Students' engagement in self-reflection has similar functions to that of scripts. By self-reflection, students analyze the learning process which leads to the learning outcome and ask themselves specific questions in order to identify their strengths and weaknesses.

Even though students' self-assessments might not be accurate, such reflection is still a pedagogically useful activity. Exploring their answers to "why" questions might enhance metacognitive awareness and result in better understanding of both the knowledge itself and the adequacy of one's own personal constructions (Eva and Regehr, 2008). In particular, these Hong Kong students were willing to figure out the reasons for the mistakes they made after receiving their marked exam papers (item 7). This is a good sign of student self-regulated learning since, as Carless and Lam (2014) argued, formative use of information provided by testing is a potentially productive way of enhancing students' performance. Such formative use is especially crucial for students in an examination-dominant context, such as that in Hong Kong.

In contrast, taking the initiative to seek feedback on their own performances appeared to be relatively more difficult for these students. Students were particularly reluctant to check whether they had fully mastered the course contents through self-assessing using reference books or notes (item 3), keep track of progress through their performance records (item 5), or asking their group members to comment on their work in group activities (item 9). With regard to students' resistance to items 3 and 5, a speculative explanation might be related to the difficulty levels of self-assessment tasks. It seems that students are used to assessing themselves with available testing questions (e.g., items 2 and 6) or past exam paper (item 4). In contrast, conducting self-assessment based on reference books/notes/performance records appears more challenging since it requires specific skills and more cognitive workload to design self-assessment tasks/questions. As for item 9, the possible reasons could be traced into the social context since socio-cultural factors have substantial influences on students' perception and experience of assessment (Carless and Lam, 2014) and, in particular, on feedback seeking behaviour (Morrison, Chen, and Salgado, 2004). As De Luque and Sommer (2000) found, in seeking feedback, collectivist cultures tended to prefer indirect inquiry and self-monitoring, but not direct-inquiry behavior since it might bring too much undesired individual attention to one-self. Mok and her colleagues (2008) further pointed out that threat to self-esteem, and maintaining face probably inhibited Chinese students from seeking help or feedback.

This study found that key demographic variables, e.g., gender and year level, played important roles in students' self-assessment practices. Girls had significantly higher self-assessment measures on both scale than did boys, meaning that girls generally used more cognitive and metacognitive learning strategies such as seeking feedback and self-reflection. This finding is consistent with those of previous studies (e.g., Bidjerano, 2005; Pokay and Blumenfeld, 1990). Gender differences in self-assessment are possibly due to differences in self-consciousness favoring females reported in literature (Alanazi, 2001; Csank and Conway, 2004). Furthermore, given that the similar nature of feedback seeking and help seeking, this finding echoed the gender difference in help seeking revealed in previous studies. For example, Cheong, Pajares, and Oberman (2004) reported that girls perceived greater benefits from help-seeking and were more willing to seek help than did boys. Nadler (1997) argued that this was probably because help-seeking had less psychological cost for girls than for boys.

The results showed that year level is another important factor related to student self-assessment practices. A general pattern is that junior students (S1 and S2) had higher measures on both scales than did their senior counterparts (S3 to S6). This is inconsistent with some previous studies (e.g., Reio, 2004; Reio and Davis, 2005) that reported that self-regulated learning ability and readiness increased with age. A closer investigation revealed that students' scores on the *Self-directed feedback seeking* scale were highest at S1 and lowest at S3, S5, and S6; students' scores on the *Self-reflection* scale were highest at S1 and lowest at S3 and S6. The most likely similarity between S3 and S6 students that might account for this finding is the pressure from high-stakes assessment. In Hong Kong, S6 students take the most important public examination, the Hong Kong Di-

ploma in Secondary Education (HKDSE). For S3 students, although there is no public examination, those students still face pressure of high-stakes internal assessment. S3 students' performances on internal assessments will determine the elective subjects they are allowed to choose at S4. This is a decision with highly important consequence as students will sit for those elective subjects in HKDSE in S6. In this sense, it seems that students are more likely to engage in self-assessment when they have less pressure from high-stakes assessment. Previous studies (e.g., Yan, 2014a, 2014b; Yan and Cheng, 2015) reported that Hong Kong teachers are facing challenges regarding how to reconcile the roles of summative and formative assessments. The present study seems to indicate that students have to deal with a similar dilemma.

The awareness of gender and year level differences in terms of self-assessment practices has implications to instructional design. Given that boys appear to be less likely to use metacognitive learning strategies, including self-directed feedback seeking and self-reflection, in their learning, teachers might need to provide more encouragement and design gender-specific supportive measures so as to enhance boys' engagement in self-assessment. Furthermore, the lowest level of engagement in self-assessment demonstrated by S3 and S6 students reminds us of the potential conflict between self-assessment and high-stakes assessment. A learning-oriented and pressure-free environment seems important in promoting student self-assessment.

There are three main lines of future research that could be explored. First, further development of the instrument is warranted, especially for the *Self-reflection* scale that has only 3 items in the current version, in order to facilitate a more thorough investigation into students' self-assessment practices. Given that students' self-assessment practices comprise different but correlated dimensions, as demonstrated in this study, a multidimensional Rasch model should be considered for future investigations so as to calibrate simultaneously both dimensions and increase the measurement precision by taking into account the attenuation of correlations between dimensions when those dimensions are calibrated separately (Wang, et al., 2006). Second, this study relied solely on data from secondary students. Further research could expand the usage these two scales to primary students and/or university students to investigate those students' self-assessment practices. Such investigations will not only provide further empirical evidences regarding the quality and appropriateness for use with primary students or university students, but also depict the characteristics of self-assessment practices of students of various ages so as to better the understanding of developmental factors and learning environment that are related to students' self-assessment practices. Third, given that students' perceptions and experiences of assessment are easily affected by socio-cultural factors (Carless and Lam, 2014), it would be worthwhile to apply the scales, with necessary modifications or additional items, on a sample from a different socio-cultural context to see whether a self-assessment practice pattern different from Hong Kong students will arise. It will shed light on contextual factors relevant to self-assessment practices from a wider socio-cultural perspective.

Acknowledgements

This research was partially supported by the Internal Research Grant (grant no. RG 72/2013-2014R) from the Hong Kong Institute of Education.

References

Alanazi, F. M. (2001). The revised self-consciousness scale: An assessment of factor structure, reliability, and gender differences in Saudi Arabia. *Social Behavior and Personality*, *29*, 763-776.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, *112*, 545-557.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*, 1-16.

Bidjerano, T. (2005, October). *Gender differences in self-regulated learning*. Paper presented at

the annual meeting of the Northeastern Educational Research Association, Kerhonkson, NY.

Blackwood, T. (2010). *Metaknowledge in higher education: Self-assessment accuracy and its association with academic achievement.* Doctoral dissertation. Available at http://nrl.northumbria.ac.uk/2233/1/blackwood.tony_dba.pdf

Bond, T. G., and Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bosma, J., Laszakovits, D., and Hattery, R. (2007). Self-assessment for maintenance of certification. *Journal of the American College of Radiology*, *4*, 45-52.

Boud, D. (1999). Avoiding the traps: Seeking good practice in the use of self assessment and reflection in professional courses. *Social Work Education*, *18*, 121-132.

Brown, G. T. L., and Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Eds.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). Thousand Oaks, CA: Sage.

Brown, G. T. L., and Harris, L. R. (2014). The future of self-assessment in classroom practice: Reframing self-assessment as a core competency. *Frontline Learning Research*, *3*, 22-30.

Carless, D., and Lam, R. (2014). Developing assessment for productive learning in Confucian-influenced settings: Potentials and challenges. In C. Wyatt-Smith, V. Klenowski, and P. Colbert, (Eds.), *Designing assessment for quality learning* (pp. 167-182). Dordrecht, Netherlands: Springer.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245-276.

Cheong, Y. F., Pajares, F., and Oberman, P. S. (2004). Motivation and academic help-seeking in high school computer science. *Computer Science Education*, *14*, 3-19.

Csank, P. A., and Conway, M. (2004). Engaging in self-reflection changes self-concept clarity: On differences between women and men, and low-and high-clarity individuals. *Sex Roles*, *50*, 469-480.

Curriculum Development Council. (2001). *Learning to Learn—The Way Forward in Curriculum.* Hong Kong: Author. Available at http://www.edb.gov.hk/en/curriculum-development/cs-curriculum-doc-report/wf-in-cur/index.html

De Luque, M. F. S., and Sommer, S. M. (2000). The impact of culture on feedback-seeking behavior: An integrated model and propositions. *Academy of Management Review*, *25*, 829-849.

Deneen, C., Brown, G. T. L., Bond, T. G., and Shroff, R. (2013). Understanding outcome-based education changes in teacher education: Evaluation of a new instrument with preliminary findings. *Asia-Pacific Journal of Teacher Education*, *41*, 441-456.

Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69-106.

Eva, K. W., and Regehr, G. (2008). "I'll never play professional football" and other fallacies of self-assessment. *Journal of Continuing Education in the Health Professions*, *28*, 14-19.

Harris, L. R., and Brown, G. T. L. (2010, May). *"My teacher's judgment matters more than mine:" Comparing teacher and student perspectives on self-assessment practices in the classroom*. Paper presented at the Annual Meeting of the American Educational Research Association, Denver, CO.

Kirby, N. F., and Downs, C. T. (2007). Self-assessment and the disadvantaged student: Potential for encouraging self-regulated learning? *Assessment and Evaluation in Higher Education*, *32*, 475-494.

Knowles, M. (1975). *Self-directed learning: A guide for learners and teachers.* San Francisco, CA: Jossey-Bass.

Kostons, D., van Gog, T., and Paas, F. (2012). Training self-assessment and task-selection

skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*, 121-132.

Kuncel, N. R., Credé, M., and Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, *75*, 63-82.

Langendyk, V. (2006). Not knowing that they do not know: Self-assessment accuracy of third-year medical students. *Medical Education*, *40*, 173-179.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*, 878.

Linacre, J. M. (2006). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2011). WINSTEPS (Version 3.73) [Computer software]. Beaverton, OR: Winsteps.com.

Mann, K., Gordon, J., and McLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education*, *14*, 595-621.

McMillan, J. H., and Hearn, J. (2008). Student self-assessment: The key to stronger student motivation and higher achievement. *Educational Horizons*, *87*, 40-49.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *Educational Measurement* (pp. 13-103). New York, NY: MacMillan.

Mok, M. M. C., Cheng, Y. C., Moore, P. J., and Kennedy, K. J. (2006). The development and validation of the self-directed learning scale (SLS). *Journal of Applied Measurement*, *7*, 418-449.

Mok, M. M. C., Kennedy, K., Moore, P., Shan, P., and Leung, S. (2008). The use of help-seeking by Chinese secondary school students: Challenging the myth of "the Chinese learner." *Evaluation and Research in Education*, *21*, 188-213.

Morrison, E. W., Chen, Y. R., and Salgado, S. R. (2004). Cultural differences in newcomer feedback seeking: A comparison of the United States and Hong Kong. *Applied Psychology*, *53*, 1-22.

Nadler, A. (1997). Autonomous and dependent help seeking: Personality characteristics and the seeking of help. In B. Sarason, I. Sarason, and R. G. Pierce, (Eds.), *Handbook of personality and social support* (pp. 258-302). New York, NY: Plenum Press.

Noonan, B., and Duncan, C. R. (2005). Peer and self-assessment in high schools. *Practical Assessment, Research and Evaluation*, *10*(17), 1-8.

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York, NY: MacGraw-Hill.

Panadero, E., and Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practical connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology*, *11*, 551-576.

Panadero, E., Alonso-Tapia, J., and Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences*, *22*, 806-813.

Panadero, E., Alonso-Tapia, J., and Reche, E. (2013). Rubrics vs. self-assessment scripts effect on self-regulation, performance and self-efficacy in pre-service teachers. *Studies in Educational Evaluation*, *39*, 125-132.

Panadero, E., Brown, G. T. L., and Strijbos, J-W. (2014, August). *The future of student self-assessment: Known unknowns and probable directions*. Paper presented at the biennial conference of the Assessment and Evaluation SIG, Madrid, Spain.

Paris, S. G., and Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, *36*, 89-101.

Pintrich, P. R., Smith, D. A., Garcia, T., and Mckeachie, W. J. (1991). *A manual for the use of*

the motivated strategies for learning questionnaire (MSLQ). Ann Arbor, MI: National Center for Research to Improve Postsecondary teaching and Learning, University of Michigan.

Pokay, P., and Blumenfeld, P. C. (1990). Predicting achievement early and late in the semester: The role of motivation and use of learning strategies. *Journal of Educational Psychology*, *82*, 41-50.

Powel, W. D., and Gray, R. (1995). Improving performance predictions by collaboration with peers and rewarding accuracy. *Child Study Journal*, *25*, 141-154.

Puustinen, M., and Pulkkinen, L. (2001). Models of self-regulated learning: A review. *Scandinavian Journal of Educational Research*, *45*, 269-286.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Reio, T. G., Jr. (2004). Prior knowledge, self-directed learning, and curiosity: Antecedents to classroom learning performance. *International Journal of Self-Directed Learning*, *1*, 18-25.

Reio, T. G., and Davis, W. (2005). Age and gender differences in self-directed learning readiness: A developmental perspective. *International Journal of Self-Directed Learning*, *2*, 40-49.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, *30*, 387-415.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research & Evaluation*, *11*, 1-13.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Wang, W. C., Yao, G., Tsai, Y. J., Wang, J. D., and Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research*, *15*, 607-620.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

Yan, Z. (2014a). Predicting teachers' intentions to implement school-based assessment using the Theory of Planned Behaviour. *Educational Research and Evaluation*, *20*, 83-97.

Yan, Z. (2014b). School-based assessment in secondary schools. In C. Marsh, and J. C-K Lee, (Eds.), *Asia's high performing education systems: The case of Hong Kong* (pp. 274-288). New York, NY: Routledge.

Yan, Z., and Cheng, E. C. K. (2015). Primary teachers' attitudes, intentions and practices regarding formative assessment. *Teaching and Teacher Education*, *45*, 128-136.

Zimmerman, B. J., and Moylan, A. R. (2009). Self-regulation: Where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, and A. C. Graesser, (Eds.), *Handbook of Metacognition in Education* (pp. 299-315). New York, NY: Routledge.

Appendix 1. Self-assessment Practice Scale

Self-directed Feedback Seeking Scale

#2 I will find some questions to test my understanding after each lesson.

#3 I often check whether I have fully mastered the course contents through assessing myself using reference books or notes.

#4 I will check whether I have fully understood course contents by doing past exam papers.

#5 I will keep track of my progress through the performance records.

#6 I will ask myself questions to check whether I have understood the course contents.

#9 I will ask my group members to comment on my work in group activities.

#10 I will invite others to test how well I have mastered the course contents.

Self-reflection Scale

#1 When the teachers ask the others, I will think about the answers.

#7 I will try to figure out the reasons for the mistakes I made after receiving the marked exam papers.

#8 I will identify areas for improvement according to the teacher's comments on my assignment.

Note: These statements are a translation, following the translation-back translation process, of the items which were administered in Cantonese.