Asia-Pacific Edu Res (2018) 27(2):123–135
https://doi.org/10.1007/s40299-018-0371-8

REGULAR ARTICLE

# The Self-assessment Practice Scale (SaPS) for Students: Development and Psychometric Studies

**Zi Yan**[1]

Published online: 2 February 2018


**Abstract** This paper describes the development and psychometric evaluation of the Self-assessment Practice Scale (SaPS), an instrument for assessing students' actions when engaged in self-assessment. Adopting a theory-driven approach, the SaPS was developed in line with the self-assessment process proposed by Yan and Brown (Assess Eval High Educ, 42(8):1247–1262, 2017). The survey was conducted with a total of 2906 Hong Kong students ranging from Primary 4 to Secondary 3. Two complementary analytical approaches, i.e., factor analysis and Rasch analysis, were applied to investigate the psychometric properties of the SaPS. Exploratory factor analysis revealed a three-factor model, while confirmatory factor analysis supported both three-factor and four-factor solutions. Rasch analysis provided further evidence of the psychometric quality of the four subscales in terms of dimensionality of the SaPS, the rating scale effectiveness, and item fit statistics. The final version of the SaPS contains 20 items in four subscales assessing students' actions in self-assessment including seeking external feedback through monitoring, seeking external feedback through inquiry, seeking internal feedback, and self-reflection.

**Keywords** Self-assessment · Self-assessment practice scale · Scale development · Rasch measurement

✉ Zi Yan
zyan@eduhk.hk

[1] D1-1/F-49, Department of Curriculum and Instruction/ Assessment Research Centre, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, NT, Hong Kong

## Introduction

Self-assessment is a core skill for self-regulated learning and life-long learning (Boud 1995; Kirby and Downs 2007; Tan 2012). Through self-assessment, learners can identify the knowledge or skills needed to learn, and they can use that awareness to motivate and guide efforts to keep abreast of new knowledge in their professions (Mok et al. 2006). Self-assessment has the potential to enhance students' commitment to, and autonomy in learning (Brown and Harris 2013) and its benefits on students' academic achievement have been well documented in the literature (e.g., Ibabe and Jauregizar 2010; Panadero and Romero 2014). However, the understanding of the inner processes of self-assessment provided in the literature remains surprisingly limited given the important role that self-assessment plays in learning. The common actions students perform when they engaging in self-assessment remain as a black box. In education or psychology, self-assessment is often referred as self-rating, i.e., giving a score or an estimate to an assessment performance students completed, or to their abilities in a particular domain (e.g., Baars et al. 2014; Kostons et al. 2010). However, self-assessment appears as a far more complex activity in the learning context. Although self-rating does have a role in learning, understanding self-assessment as a sophisticated process would enhance its standing in improving learning.

## Theoretical Framework

Self-assessment is not only about evaluating one's own work against selected criteria, but also about students' efforts in drawing on their diverse personal resources to facilitate such evaluations (Davis et al. 2006; Ryan 2014; Yan and Brown 2017). This implies that the self-

assessment process involves various actions. For example, one of the important actions is likely to be seeking feedback, taking the initiative to gather feedback about the quality of their work from various sources for assessing their own performance. However, such feedback itself does not necessarily lead to a self-assessment judgment that is meaningful to learning. Students have to reflect upon their learning—with the support of feedback—so as to develop self-assessment from mere self-rating to far more comprehensive understanding of their strengths and weaknesses, how to improve, and so forth.

Following this line of reasoning, Yan and Brown (2017) defined self-assessment as "a process during which students collect information about their own performance, evaluate and reflect on the quality of their learning process and outcomes according to selected criteria, to identify their own strengths and weaknesses." (p. 2). They also proposed a cyclical self-assessment process consisting of three common and sequenced actions including *determining the performance criteria*, *self-directed feedback seeking*, and *self-reflection*. When engaging in self-assessment, students first determine the assessment criteria against which their own performance is assessed. They then take the initiative to seek feedback regarding their learning from various sources which could be classified into two major categories, i.e., external and internal sources. External feedback could be obtained through monitoring (e.g., doing extra exercises or past test papers), and/or through inquiry with people (e.g., teachers, peers). In contrast, internal feedback refers to internally generated reactions to their own performance, such as emotions, physical sensation, and internal states. With the support of relevant feedback, students reflect on the quality of their own performance, and identify their own strengths and weaknesses. Based on such self-reflection, a self-assessment judgement is then arrived at and this judgement is subjected to continuous reconsideration based on new sources of feedback or different assessment criteria.

## Extant Instruments for Self-assessment Practices

Compared with studies investigating the effects of self-assessment on student learning, the research focusing on self-assessment practice itself remains limited in the literature. This could be due to a research gap pointed out in a recent review (Panadero et al. 2016) that there is a lack of common understanding of self-assessment inner processes and more studies should be conducted to obtain detailed data about the self-assessment practices using solid instruments. A better understanding and assessment of students' inner processes, i.e., the actions students incorporated in self-assessment practices, can contribute both to enhancing student self-assessment, and, optimizing its potential effects on learning. This would help to address important research/learning questions, as: what factors can encourage or impede student self-assessment? Where and when, in the self-assessment process, do the positive effects of self-assessment impact on learning? How should we design instruction or intervention programmes to promote students' self-assessment practices and maximize their positive effects? To the author's knowledge, there is only one instrument (i.e., Yan 2016a), that is specifically designed for assessing different actions in the self-assessment process. Apart from that, some elements of self-assessment practice can be found in instruments assessing related constructs, such as feedback seeking (e.g., Ashford 1986; Hwang and Arbaugh 2006; Krasman 2010; Swann et al. 1989). Some instruments assessing self-regulated learning (e.g., Cleary 2006; Mok et al. 2006; Pintrich et al. 1991; Suh et al. 2015) also contain subscales on self-monitoring or self-reflection. However, these instruments either describe self-assessment in a general way, such as "I evaluate my performance in learning", or focus on only a particular action in student self-assessment and, therefore, fail to provide a comprehensive picture of the whole self-assessment process. Furthermore, none of these instruments were developed under a theoretical model about self-assessment process, and the key component of internal feedback is missing from these self-reports.

Yan (2016a) developed a preliminary 10-item scale to assess secondary students' self-assessment practices. The scale contains two subscales with seven items assessing self-directed feedback seeking and three items gauging self-reflection. Although this scale was in line with the self-assessment process model proposed by Yan and Brown (2017), it should be open to further improvement to provide a more comprehensive picture of students' self-assessment practices. For instance, the self-directed feedback seeking subscale did not include items about internal feedback. The items for external feedback were not further classified into sub-categories (i.e., monitoring vs. inquiry) and were not sufficiently comprehensive with some important feedback sources, such as parents and teachers, omitted. Furthermore, having only three items, the self-reflection subscale results in relatively low Rasch person reliability. Adding more items to this subscale would be likely to increase the reliability and provide a more holistic depiction of students' self-reflection activities during self-assessment.

Considering the importance of self-assessment in student learning and the lack of comprehensive, validated, and psychometrically sound instrument for assessing various components of self-assessment practices, the purpose of this study was to develop an instrument to assess student self-assessment practice. This article describes the

development of the instrument and provides preliminary evidence of the psychometric properties of the ensuing scale.

## Method

### Scale Development

Adopting a theory-driven approach, the Self-assessment Practice Scale (SaPS) was developed in line with the self-assessment process proposed by Yan and Brown (2017). In this framework, the self-assessment process contains three actions: determining the performance criteria, self-directed feedback seeking, and self-reflection. From the perspective of scale development, it is not straightforward to develop a scale for determining performance criteria, because students tend to select only one self-assessment criterion in most scenarios. It is true that students may change the performance criterion even within the same self-assessment task, but that will occur in another round of self-assessment once the criterion is changed. Consequently, the SaPS tends to focus on the remaining key actions, i.e., self-directed feedback seeking and self-reflection. These apply to all self-assessment scenarios, and are independent of any performance criteria selected by students. Self-directed feedback seeking refers to the process by which students initiate and take responsibility for seeking feedback from internal and external sources for the purpose of self-assessment. Internal feedback comes from "the self", such as emotions, feelings, and subjectively experienced internal states; while external feedback comes from "the outside" which could be further divided into those from objective external evidence against legitimate standards (e.g., past tests and reference books) and subjective evidence from people (e.g., teachers, peers, and parents). According to the terminology of Ashford and Cummings (1983), they could be named "feedback seeking through monitoring" and "feedback seeking through inquiry", respectively. Self-reflection refers to the action by which students go back and reflect upon their learning processes and outcomes with the support of available/gathered feedback to identify their own strengths and weaknesses. Then, theoretically, the SaPS could have four subscales, namely, seeking external feedback through monitoring (SEFM), seeking external feedback through inquiry (SEFI), seeking internal feedback (SIF), and self-reflection (SR).

Since the scale questions were to be completed by Hong Kong students, items were developed in Cantonese. An initial item pool of 51 items was generated from an in-depth literature review, focus group interviews with five teachers, and 12 primary and 12 secondary students. The initial item pool was then reduced to 31 items by removing redundant and/or less relevant items. The resultant 31 items were reviewed by a panel of five experts, three from educational assessment and two experienced teachers, to examine the evidence for face and content validity. The five experts responded to the following question for each item: "Is the item 'essential', 'useful, but not essential', or 'not necessary' to assess student self-assessment practice?" The content validity ratios[1] (CVRs; Lawshe 1975) were computed based on experts' responses. Only the items with positive CVR indicators (i.e., three or more experts indicating essential) were retained. The expert panel was also asked to comment on the readability of the items and appropriate modifications were made. Nine items with negative CVR were discussed, then removed, and a resultant 22-item survey was generated with 6 SEFM items, 4 SEFI items, 5 SIF items, and 7 items for SR. A common six-point Likert-type response scale (1 = *Strongly disagree*, 2 = *Disagree*, 3 = *Slightly disagree*, 4 = *Slightly agree*, 5 = *Agree*, 6 = *Strongly agree*) was adopted.

**Table 1** Details of the two samples

| | Sample 1 | | Sample 2 | |
|---|---|---|---|---|
| | *N* | % | *N* | % |
| Male | 775 | 52.0 | 707 | 49.9 |
| Female | 708 | 47.5 | 703 | 49.6 |
| Primary 4 | 203 | 13.6 | 185 | 13.1 |
| Primary 5 | 246 | 16.5 | 211 | 14.9 |
| Primary 6 | 240 | 16.1 | 232 | 16.4 |
| Secondary 1 | 260 | 17.4 | 254 | 17.9 |
| Secondary 2 | 229 | 15.4 | 237 | 16.7 |
| Secondary 3 | 312 | 20.9 | 297 | 21.0 |
| Overall | 1490 | 100 | 1416 | 100 |

### Sample

The survey was conducted with a convenience sample of 2906 Hong Kong students from 18 primary schools and 11 secondary schools. Participating students ranged from Primary 4 to Secondary 3 (aged 9–14 years). Two questionnaire packages (A and B), each containing the SaPS and a different set of other instruments, were randomly assigned to participating students in the survey. The current study used only the data collected with the SaPS and this resulted in two samples of respondents. The data from students who completed package A (Sample 1) were used for exploratory factor analysis (EFA) and the data from students completing package B (Sample 2) were for

[1] CVR = $(n_e - N/2)/(N/2)$, where $n_e$ = number of experts indicating "essential", $N$ =total number of experts.

confirmatory factor analysis (CFA) and Rasch analysis. Table 1 presents the details of the two samples.

### Data Analysis

Two complementary analytical approaches, i.e., factor analysis followed by Rasch analysis (Rasch 1960), were applied to investigate the psychometric properties of the SaPS. Although these two analytical methods are based on quite different principles, applying both to the same set of data might provide a more circumspect interpretation of research results and provoke discussions regarding the relative advantages of each approach (Richardson 2005). A number of empirical studies (e.g., Chang and Engelhard 2016; Deneen et al. 2013; Hart et al. 2013; Primi et al. 2014; Yan 2016a) have demonstrated that both factor analysis and Rasch analysis can each provide unique information and, therefore, might bring benefits to validating and refining instruments.

EFA (SPSS version 24) using principal axis factoring method and promax rotation was conducted on the data from Sample 1 in an attempt to identify the factor groupings underlying the items. Promax rotation method was chosen because, in view of the underlying theory, the factors were allowed to be correlated with one another. Three criteria, including the eigenvalue test, scree plots, and interpretability of the rotated factors, were used to determine the number of factors in the data matrix. A factor loading higher than 0.4, with no cross-loading above 0.3, was used as the item selection criteria for each factor.

CFA (AMOS version 24) was conducted on the data from Sample 2 with the aim of testing the fit between those empirical data and the hypothesized factor structure derived from the results of the EFA on Sample 1. Multiple fit indices were used to examine the model-data fit including Chi square ($\chi^2$), the goodness-of-fit index (GFI), the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). Values of GFI and CFI greater than 0.90, and SRMR and RMSEA values less than 0.08 (Hu and Bentler 1999; McDonald and Ho 2002) were considered to indicate acceptable fit.

Subsequently, Rasch analysis was employed on the data from Sample 2 to provide further detailed psychometric evidence of the properties of the SaPS. Rasch analysis examines scale quality by checking the extent to which items that form a scale reflect a single underlying unidimensional latent construct. To compute the unidimensionality metrics, the Rasch model implements a "data fit the model" approach that requires the empirical data to satisfy a priori requirements essential for the purpose of fundamental measurement (Andrich 2004; Bond and Fox 2015). Given that self-assessment practice is considered to be a multi-faceted construct that contains sets of different but related actions, a multidimensional Rasch-based model (Adams et al. 1997), rather than the conventional unidimensional Rasch model, was employed with these data. In a multidimensional Rasch-based model, all subscales are calibrated simultaneously, in order that they can be examined individually. Measurement precision on each subscale is enhanced by taking into account the correlations between the subscales in the computations. The criteria used to examine the model-data fit include response category functioning and item fit statistics (i.e., Infit MNSQ and Outfit MNSQ). Infit/Outfit MNSQ falling into the range between 0.75 and 1.33 are routinely regarded as indicating sufficient fit to the Rasch model (Wilson 2005). Furthermore, since key demographic variables including gender and year level were found to have substantial impact on self-assessment practices (Yan 2016b), the evidence for differential item functioning (DIF) across gender and year level was also checked. DIF analysis checks for construct equivalence across groups. The existence of DIF has a potentially detrimental impact on scale quality, since different respondent groups might have different interpretations of, or perspectives on the items. DIF is indicated by the difference of item difficulty across groups after controlling the levels of latent trait; a difference equal to or larger than 0.5 logits was considered as indicating substantive DIF (Wang et al. 2006).

The present study examined the validity of the SaPS according to Messick's (1995) framework. That is, validity should be viewed as a unified concept which could be examined from six distinguishable aspects including: content, substantive, structural, generalizability, external, and consequential aspects of validity (Messick 1995). In particular, the content aspect of validity (i.e., the instrument's relevance, representativeness, and clarity) was ensured by the theory-driven procedure of scale development and expert review. The substantive aspect of validity was satisfied by a valid theoretical self-assessment process model, i.e., Yan and Brown (2017), complemented by considering evidence from the response category functioning, the data-to-model fit statistics, and the empirical hierarchy of item measures in Rasch analysis. The focus of this aspect is the theory of the construct and whether the items are reflective of the theory. The structural aspect of validity refers to the consistency between items' empirical relations to one another and the theoretical specifications. This was investigated by EFA and CFA. The evidence for the generalizability aspect of validity could be drawn from the results of DIF across gender and year level. The DIF results could determine whether the instrument works invariantly across gender and year level groups. However, the present study would not provide evidence for the external and consequential aspects of validity. In addition, reliability will be

checked through Cronbach's α estimates and Rasch reliabilities.

## Results

### Exploratory Factor Analysis

The development of the SaPS was a theory-driven process. Even though EFA is routinely used when there is no theorized factor structure underlying items (Thompson 2004), it might be profitable to explore what the data will reveal as the initial step of the validation without the precursor of the theoretical constraint. Therefore, EFA was conducted with the data from Sample 1 to identify items reflecting factors of self-assessment. The skewness and kurtosis values were computed for each item to examine the normality of their distributions. The skewness index ranged from − 0.92 to − 0.09 and kurtosis index from -0.93 to 0.86, implying that the data were approximately normally distributed (Kline 2010). The values for Kaiser–Meyer–Olkin measure of sampling adequacy was 0.96 and Bartlett's test of sphericity was $\chi^2(231) = 14{,}794.64$, $p < 0.001$, indicating that the data were suitable for factor analysis. The corrected item-total correlations for each item were also computed (see Table 2). Three factors with eigenvalues higher than 1 which could be meaningfully interpreted were identified. The scree test (Cattell 1966) also supported the three-factor solution. The factors were identified, from the theoretically driven item content, as seeking external feedback, seeking internal feedback, and self-reflection. This result was slightly different from that posited from the original theoretical model, as described in Yan and Brown (2017) and the scale development section above, in that SEFM items and SEFI items loaded onto one factor. Only one item—supposed to be about seeking external feedback (*Item #5: I compare my performance against that of the best students in the class*)—did not load at the 0.40 level on any factor. One item written to assess seeking internal feedback (*Item #15: I compare my performance to my own expectations to see if I am doing a good job or not*) cross-loaded heavily (0.48) on the self-reflection factor. Apart from that all items in SIF and SR loaded on factors in line with the original theoretical model. The rotated factor structure of the scale is presented in Table 2.

**Table 2** Rotated factor structure of the SaPS

| Item | Corrected item-total correlation | Factor | | |
|---|---|---|---|---|
| | | Seeking external feedback | Self-reflection | Seeking internal feedback |
| Item 1 | 0.66 | **0.516** | 0.274 | − 0.159 |
| Item 2 | 0.68 | **0.497** | 0.292 | − 0.151 |
| Item 3 | 0.62 | **0.617** | 0.111 | − 0.055 |
| Item 4 | 0.67 | **0.443** | 0.320 | − 0.018 |
| Item 5 | 0.50 | 0.275 | − 0.004 | 0.279 |
| Item 6 | 0.63 | **0.526** | 0.069 | 0.118 |
| Item 7 | 0.63 | **0.882** | − 0.239 | 0.086 |
| Item 8 | 0.62 | **0.690** | 0.029 | − 0.065 |
| Item 9 | 0.65 | **0.631** | 0.062 | 0.058 |
| Item 10 | 0.67 | **0.720** | − 0.056 | 0.068 |
| Item 11 | 0.58 | 0.034 | 0.282 | **0.476** |
| Item 12 | 0.41 | − 0.051 | − 0.036 | **0.562** |
| Item 13 | 0.52 | 0.085 | − 0.100 | **0.789** |
| Item 14 | 0.46 | − 0.074 | 0.064 | **0.718** |
| Item 15 | 0.63 | − 0.025 | **0.483** | 0.**426** |
| Item 16 | 0.65 | 0.053 | **0.728** | − 0.037 |
| Item 17 | 0.66 | 0.229 | **0.467** | 0.058 |
| Item 18 | 0.66 | − 0.016 | **0.764** | 0.008 |
| Item 19 | 0.70 | 0.093 | **0.696** | 0.003 |
| Item 20 | 0.69 | 0.094 | **0.774** | − 0.073 |
| Item 21 | 0.61 | − 0.180 | **0.857** | 0.023 |
| Item 22 | 0.67 | 0.168 | **0.556** | 0.065 |

Based on the EFA results, seeking external feedback (SEF) had 9 items, seeking internal feedback (SIF) contained 4 items, and self-reflection (SR) had 7 items. The three-factor solution accounted for 51% of the common variance.

### Confirmatory Factor Analysis

CFA was conducted on the data from Sample 2 to cross validate the factor structure of the scale identified by EFA. In addition to the three-factor model (Model 1, see Fig. 1) identified by EFA, a rival model (Model 2, see Fig. 2), which was consistent with the original hypothesized model proposed by Yan and Brown (2017), was also tested to confirm if one model was more appropriate. The rival model (Model 2) was a high-order factor model with four factors representing the four actions, including SEFM, SEFI, SIF, and SR, that students would engage in self-assessment. SEFM and SEFI contributed to a second-order factor, namely, seeking external feedback (SEF), which, together with SIF, belonged to the factor seeking feedback (SF). All items were assigned to the underlying factor in harmony with the allocations predicted by the theory. Given that Chi-square statistics are easily affected by samples size, especially when sample sizes are much over 200 (Newsom 2012), and the information criteria (e.g., Akaike's Information Criterion, AIC, and Bayesian Information Criterion, BIC) have better performance than Chi squares in model selection (Lin 2006), AIC and BIC were used for model comparison. According to Kline (2010), a smaller AIC or BIC indicates a more parsimonious model.

It can be seen from Table 3 that both models fit the data well, with Model 2 having slightly better fit statistics and smaller AIC and BIC. The results showed that the four-factor solution in Model 2 accounted for the data better than did Model 1.

Figures 1 and 2 display the standardized factor loadings and correlations for the two models on the data from Sample 2. The factor loadings were acceptable for all factors. For Model 1, the factor loadings ranged from 0.66 to 0.73 for SEF, 0.62 to 0.81 for SIF, and 0.67 to 0.81 for SR. For Model 2, the factor loadings ranged from 0.65 to 0.78 for SEFM, 0.74 to 0.77 for SEFI, 0.62 to 0.81 for SIF, and 0.67 to 0.81 for SR.

### Rasch Analysis

A multidimensional Rasch Rating Scale model using ConQuest version 2.0 software (Wu et al. 2007) was applied to the data from Sample 2. Given that the theoretically consistent model (Model 2) demonstrated better fit than did Model 1 that was generated by EFA, the multidimensional Rasch analysis calibrated the four subsets of items, as specified in Model 2, of the SaPS simultaneously. The category functioning of the six-point rating scale response options was first examined. The step calibrations, the calibrated measures of the transitions between adjacent categories, should increase monotonically if all categories function well. The ideal step distances should be between 1.4 logits and 5.0 logits. For a rating scale with four or more categories, shorter step distances are acceptable (Linacre 2002). The results showed that the step calibrations increased monotonically from − 1.28, − 1.14, − 0.78, 0.87, to 2.34 logits, indicating that the rating scale was acceptable, although it would be better if the step distances between the first three step calibrations were larger. The correlations among the four subscales (see Table 4) ranged from 0.45 to 0.71. This result further justified the use of multidimensional Rasch analysis, since inter-subscale correlations were neither too low (indicating lack of relationship), nor too high (indicating redundancy) (Bond and Fox 2015).

The model-data fit was examined by item fit statistics (i.e., Infit and Outfit MNSQ) of each item. In ConQuest analyses, the items are being analysed in one of the four subscales, so that fit statistics are for items to subscale—not to the whole instrument. An initial ConQuest analysis identified one misfitting item in SEFM (Infit MNSQ = 1.67; Outfit MNSQ = 1.81) (*I compare my performance against that of the best students in the class*) and one marginally misfitting item in SIF (Infit MNSQ = 1.35; Outfit MNSQ = 1.27) (*I compare my performance to my own expectations to see if I am doing a good job or not*). This result was in line with the EFA results. These two items were, therefore, removed and the data were reanalyzed. The remaining 20 items demonstrated sufficient fit to the model, revealing that all items were assessing the latent trait as the theory hypothesized. The DIF analyses across gender and year level revealed no substantive DIF by gender for all items. Only one item from SEFI scale (Item 7: *I ask my family members to give me advice on my work*) demonstrated DIF across year level. The item difficulties, their associated standard errors, item fit statistics, and DIF results are presented in Table 5.

The hierarchy of item and student measures is displayed in the ConQuest Wright map of the four component variables (Fig. 3). The Rasch model calibrates person measures and item difficulties along the same latent trait scale. The four continua on the left-hand side show the distribution of students in the four subscales of self-assessment. The students with higher levels in self-assessment are placed at the top of the continuum and those with lower levels in self-assessment are placed at the bottom. On the right-hand side of the scale, the items are grouped in the four subscales. Items with higher difficulty level (less likely to be

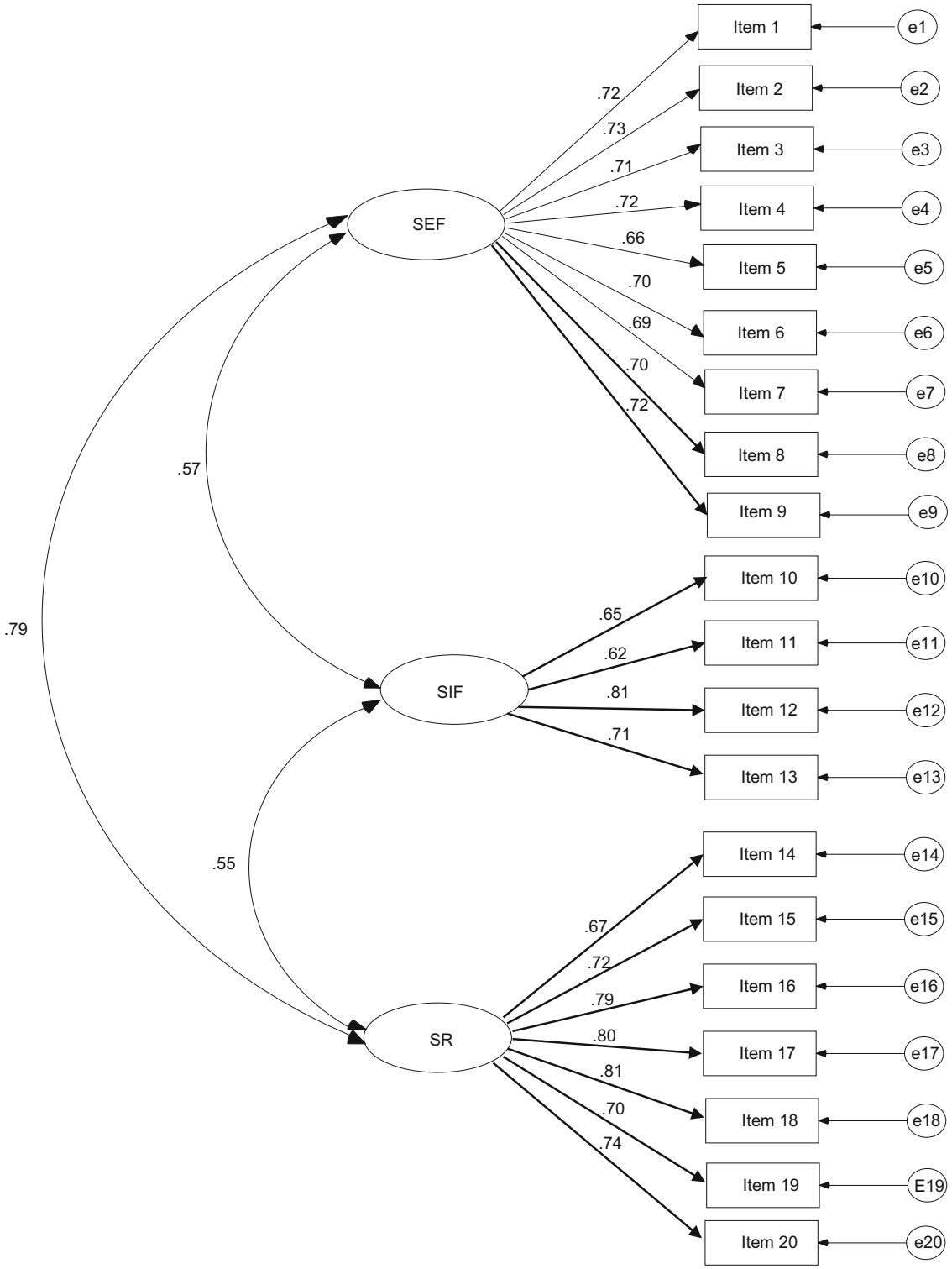


**Fig. 1** Model 1 with standardized factor loadings on seeking external feedback (SEF), seeking internal feedback (SIF), and self-reflection (SR) and their correlations

endorsed) are placed at the top and the items with lower difficulty level (more likely to be endorsed) are placed at the bottom. In addition, the rating scale with six categories is indicated by the last column. It can be seen from Table 5 and Fig. 3 that the item difficulties were in the range of − 0.43 to 0.16 logits. The most difficult item for students

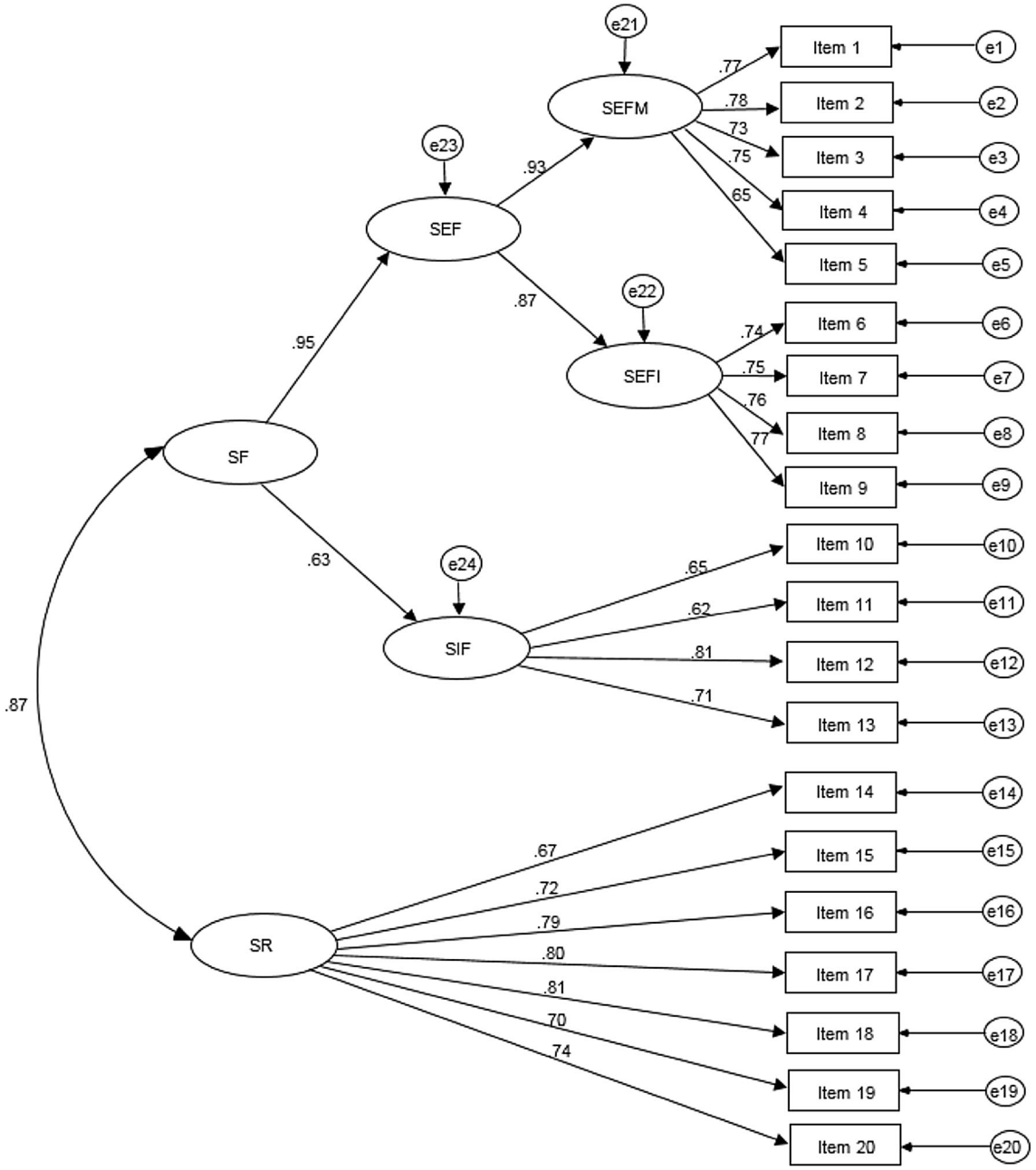


**Fig. 2** Model 2 with standardized factor loadings on seeking external feedback through monitoring (SEFM), seeking external feedback through inquiry (SEFI), seeking internal feedback (SIF) and self-reflection (SR), and their correlations

**Table 3** CFA goodness-of-fit indices for the two models on Sample 2

| | $\chi^2$ | df | GFI | CFI | RMSEA | SRMR | AIC | BIC |
|---|---|---|---|---|---|---|---|---|
| Model 1 | 1391.6, $p < 0.001$ | 167 | 0.90 | 0.92 | 0.07 | 0.05 | 1477.59 | 1703.58 |
| Model 2 | 1031.9, $p < 0.001$ | 165 | 0.93 | 0.94 | 0.06 | 0.05 | 1121.87 | 1358.37 |

to endorse was SEFM Item #3: (*I keep track of my progress by recording my performance*). The least difficult item was SR Item#19: (*I pay attention to my assessment results to identify what I can do better next time*). The Wright map shows that the ranges of item difficulty on the four subscales were quite small compared to the ranges of student

**Table 4** Means, standard deviations, and correlations between the four subscales and the overall scale

| | Mean | SD | SEFM | SEFI | SIF | SR | Overall |
|---|---|---|---|---|---|---|---|
| SEFM | 4.21 | 0.98 | – | | | | |
| SEFI | 3.90 | 1.15 | 0.71 | – | | | |
| SIF | 4.29 | 0.99 | 0.45 | 0.46 | – | | |
| SR | 4.39 | 0.97 | 0.69 | 0.63 | 0.49 | – | |
| Overall | 4.23 | 0.84 | 0.87 | 0.84 | 0.69 | 0.89 | – |

ability. However, the items share a six-point rating scale, as presented by the last column, and this represented a reasonably wide range of difficulty on the logit scale which corresponds more closely to the range of student ability revealed by this survey.

## Reliability

The SaPS demonstrated good reliability in the studied population. Cronbach's $\alpha$ for the four subscales were SEFM 0.85, SEFI 0.84, SIF 0.79, and SR 0.90, respectively, indicating satisfactory levels of internal consistency. Furthermore, the Rasch reliabilities were similarly satisfactory. The corresponding EAP/PV reliabilities generated by ConQuest for the four sub-sets of items were 0.88, 0.88, 0.80, and 0.90, respectively.

## Discussion

The present study aimed at developing and validating an instrument for measuring student self-assessment practice using two samples of Hong Kong students. Collectively, the results of both factor analysis and Rasch analysis were supportive of the SaPS as an acceptable measure for use

**Table 5** Item statistics in Rasch analysis

| Scale/item | Item measure[a] | SE | Infit MNSQ | Outfit MNSQ | DIF[b] | |
|---|---|---|---|---|---|---|
| | | | | | Gender | Year level |
| Seeking external feedback monitoring (SEFM) | | | | | | |
| Item 1 | − 0.04 | 0.02 | 0.98 | 0.96 | 0.05 | 0.50 |
| Item 2 | − 0.12 | 0.02 | 0.88 | 0.84 | 0.15 | 0.14 |
| Item 3 | 0.16 | 0.02 | 0.99 | 0.99 | 0.06 | 0.19 |
| Item 4 | − 0.05 | 0.02 | 0.86 | 0.9 | 0.08 | 0.24 |
| Item 5 | 0.05 | 0.04 | 1.16 | 1.18 | 0.06 | 0.35 |
| Seeking external feedback inquiry (SEFI) | | | | | | |
| Item 6 | 0.12 | 0.02 | 1.04 | 1.07 | 0.24 | 0.31 |
| Item 7 | 0.03 | 0.02 | 1.09 | 1.07 | 0.01 | *0.65* |
| Item 8 | − 0.11 | 0.02 | 1.03 | 1.01 | 0.20 | 0.42 |
| Item 9 | − 0.04 | 0.04 | 0.96 | 0.96 | 0.03 | 0.38 |
| Seeking internal feedback (SIF) | | | | | | |
| Item 10 | − 0.02 | 0.02 | 1.15 | 1.15 | 0.10 | 0.14 |
| Item 11 | − 0.04 | 0.02 | 1.25 | 1.24 | 0.05 | 0.13 |
| Item 12 | 0.09 | 0.02 | 0.89 | 0.89 | 0.07 | 0.14 |
| Item 13 | − 0.03 | 0.04 | 1.13 | 1.1 | 0.01 | 0.13 |
| Self-reflection (SR) | | | | | | |
| Item 14 | 0.09 | 0.02 | 1.17 | 1.2 | 0.08 | 0.27 |
| Item 15 | 0.10 | 0.02 | 1.15 | 1.17 | 0.03 | 0.31 |
| Item 16 | 0.09 | 0.02 | 0.91 | 0.91 | 0.03 | 0.25 |
| Item 17 | 0.12 | 0.02 | 0.82 | 0.85 | 0.02 | 0.17 |
| Item 18 | 0.06 | 0.02 | 0.77 | 0.8 | 0.01 | 0.17 |
| Item 19 | − 0.43 | 0.02 | 1.05 | 1.01 | 0.09 | 0.15 |
| Item 20 | − 0.03 | 0.06 | 0.94 | 0.94 | 0.01 | 0.31 |

[a]All measures are in logits

[b]The figures for gender DIF are the absolute values in logits of item difficulty differences between males and females. The figures for year level DIF are the absolute values in logits of the largest item difficulty differences among different year levels

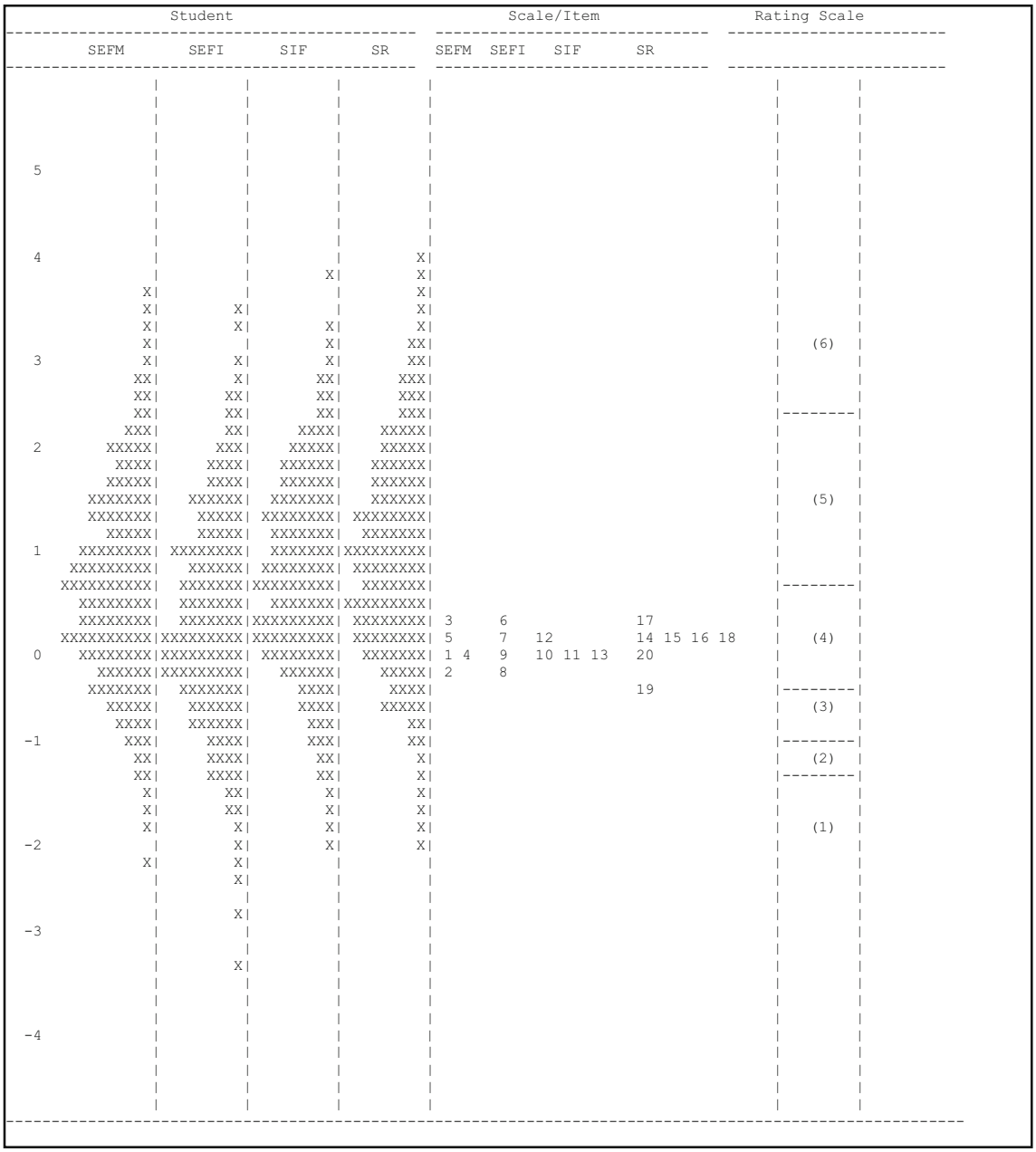


**Fig. 3** Wright map for the 20-item SaPS on four subscales: SEFM, SEFI, SIF, and SR. Each "X" represents ten cases

with Primary 4 to Secondary 3 students in Hong Kong. The final version of the SaPS comprises of 20 items accessing four common actions in student self-assessment process: SEFM (5 items), SEFI (4 items), SIF (4 items), and SR (7 items). An English version of the scale generated following back translation is shown in "Appendix".

The EFA of the first data set revealed a three-factor model. The CFA of the second data set confirmed that three-factor structure. However, the results of CFA found that a theoretically consistent model with four factors—by dividing the factor seeking external feedback into its two theorized factors, namely, SEFM and SEFI—had better model-data fit than did the three-factor solution identified by the EFA.

Rasch analysis further examined the dimensionality of the four subscales and provided useful item-level information regarding the psychometric properties of the scale. The findings supported that all the 20 items in the final version of the SaPS fit to the Rasch model, i.e., they are measuring in ways that are consistent with the underlying theory. The six-point rating scale was acceptable in its current form, since the step calibrations were ordered. However, the first three steps were not optimally spread according to Linacre's (2002) recommendations. Future

studies using the SaPS could trial a four-point rating scale. The results also indicated that the SaPS item measures were invariant (within error) across gender and year level; all items, except one, had no substantive DIF. The only item with DIF is about seeking external feedback through inquiry (*I ask my family members to give me advice on my work*). Students from different year levels seem to have different interpretations on what it means to "ask for feedback from family member" and respond in different manners. This should be further tested in future studies and responses to this item are marked for possible exclusion from across-year-level comparisons if the DIF remains in future large-scale studies.

The Rasch analysis has identified areas for further improvement. For example, it revealed that the four subscales cover only a relatively small range of item difficulties along the latent trait scale. Additional items could be considered so as to capture a wider range of difficulty levels and to accommodate students with relatively higher and lower levels of self-assessment practice.

Notwithstanding some areas requiring further development, the SaPS provides a viable option for researchers who are interested in investigating students' self-assessment practices and related topics. The uniqueness of this instrument is that it is theory driven in line with the process of self-assessment proposed by Yan and Brown (2017). Researchers can use the SaPS to collect data that are crucial for constructing a comprehensive measure about the actions students engaged in self-assessment practices. A better understanding and assessment of students' inner processes of self-assessment can inform teaching and learning in terms of promoting student self-assessment and optimizing its potential effects on learning.

Two limitations are associated with the present study. First, the sample used came from one culture, i.e., Hong Kong where Cantonese is spoken, and Confucianism is the dominant culture. Self-assessment practices are highly likely to be influenced by social interactions and cultural values. To validate the scale further, the next step would be to trial the scale with samples from slightly different sociocultural contexts, such as in Macau and mainland China, then in Western cultures, and examine the invariance of the SaPS items across cultures. Second, the present study offered initial evidence of the content, substantive, structural, and generalizability aspects of validity for the SaPS, but could not examine the external and consequential aspects of validity. Future studies should work on this direction to provide further credence to assessing self-assessment practice using the SaPS.

In summary, the present study has recognized the significance of investigating student self-assessment practice and supported the theoretical framework of the self-assessment process model proposed by Yan and Brown (2017). The development of the SaPS provides a useful tool for measuring various self-assessment actions in primary and secondary students. The information provided by this instrument can contribute to a better understanding of students' self-assessment practices and, therefore, inform teachers and researchers about how to enhance students' self-assessment as well as their self-regulated learning.

**Acknowledgements** I am grateful to Professor Gavin T. L. Brown for his constructive comments on drafting and revising the SaPS items. Special thanks also go to Professor LEE Chi Kin John, Dr. KO Po Yuk, and the team of the project "Fostering Communities of Practice for Effective Teaching and Learning" for their assistance in data collection.

**Author's Contribution** Dr. Zi Yan is an associate professor at The Education University of Hong Kong, and Associate Head of the Department of Curriculum and Instruction. His research interests focus on Rasch measurement, assessment in school and higher education contexts, with an emphasis on student self-assessment and self-regulated learning.

**Funding** This work was supported by the General Research Fund (GRF) (Project Number: 18605715) of the Research Grants Council of Hong Kong.

## Appendix

### Self-assessment Practice Scale (SaPS)

*Seeking External Feedback Through Monitoring (SEFM)*

1. I check whether I have mastered the course content by doing extra exercises.
2. I check whether I have fully understood the course content by doing past exam papers.
3. I keep track of my progress by recording my performance.
4. I ask myself questions in my head to check whether I have understood the course content.
5. I check my performance against the answers in the text book or on a website.

*Seeking External Feedback Through Inquiry (SEFI)*

6. I ask my teachers to give me feedback about my performance.
7. I ask my family members to give me advice on my work.
8. I ask my friends to tell me how to improve my learning.

9. I ask my fellow group members to evaluate my contributions to group work tasks.

*Seeking Internal Feedback (SIF)*

10. My gut feelings tell me whether my work is good or bad.
11. My emotions influence my evaluation on my learning performance.
12. How my body feels tells me how well I am doing.
13. My intuition tells me if I am doing a good job or not.

*Self-reflection (SR)*

14. I seek out the reasons for mistakes I made after getting back marked work.
15. I think about how much sense the comments of other people (e.g., teachers, family members, and friends) regarding my work make to me.
16. Any areas I am unsure of after finishing my work, I go over again.
17. As I study, I think about whether the way I am studying is really helping me learn.
18. When I do exercise, I look at what I got wrong or did poorly on to guide me as to what I should learn next.
19. I pay attention to my assessment results to identify what I can do better next time.
20. I reflect on my weaknesses when I discuss study-related issues with my classmates.

## References


Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23.

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

Ashford, S. J. (1986). Feedback-seeking in individual adaptation: A resource perspective. *Academy of Management Journal, 29*(3), 465–487.

Ashford, S. J., & Cummings, L. L. (1983). Feedback as an individual resource: Personal strategies of creating information. *Organizational Behavior and Human Performance, 32*(3), 370–398.

Baars, M., Vink, S., van Gog, T., de Bruin, A., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction, 33,* 92–107.

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

Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.

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

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

Chang, M. L., & Engelhard, G. (2016). Examining the teachers' sense of efficacy scale at the item level with Rasch measurement model. *Journal of Psychoeducational Assessment, 34*(2), 177–191.

Cleary, T. J. (2006). The development and validation of the self-regulation strategy inventory-self-report. *Journal of School Psychology, 44,* 307–322.

Davis, D. A., Mazmanian, P. E., Fordis, M., van Harrison, R., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence. *Journal of American Medical Association, 296*(9), 1094–1102.

Deneen, C., Brown, G. T. L., Bond, T. G., & 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.

Hart, C. O., Mueller, C. E., Royal, K. D., & Jones, M. H. (2013). Achievement goal validation among African American high school students: CFA and Rasch results. *Journal of Psychoeducational Assessment, 31*(3), 284–299.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Hwang, A., & Arbaugh, J. B. (2006). Virtual and traditional feedback-seeking behaviors: Underlying competitive attitudes and consequent grade performance. *Decision Sciences Journal of Innovative Education, 4*(1), 1–28.

Ibabe, I., & Jauregizar, J. (2010). Online self-assessment with feedback and metacognitive knowledge. *Higher Education, 59,* 243–258.

Kirby, N. F., & Downs, C. T. (2007). Self-assessment and the disadvantaged student: Potential for encouraging self-regulated learning? *Assessment & Evaluation in Higher Education, 32*(4), 475–494.

Kline, R. B. (2010). *Principles and practice of structural equation modeling* (3rd ed.). New York: The Guilford Press.

Kostons, D., van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education, 54*(4), 932–940.

Krasman, J. (2010). The feedback-seeking personality: Big five and feedback-seeking behavior. *Journal of Leadership & Organizational Studies, 17*(1), 18–32.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28,* 563–575.

Lin, T. H. (2006). A comparison of model selection indices for nested latent class models. *Monte Carlo Methods and Applications, 12*(3), 239–259.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85–106.

McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64–82.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50,* 741–749.

Mok, M. M. C., Cheng, Y. C., Moore, P. J., & Kennedy, K. J. (2006). The development and validation of the self-directed learning scale (SLS). *Journal of Applied Measurement, 7*(4), 418–449.

Newsom, J. T. (2012). Some clarifications and recommendations on fit indices. Retrieved from http://www.upa.pdx.edu/IOA/newsom/semclass/ho_fit.pdf.

Panadero, E., Brown, G. T., & Strijbos, J. W. (2016). The future of student self-assessment: A review of known unknowns and potential directions. *Educational Psychology Review, 28,* 803–830.

Panadero, E., & Romero, M. (2014). To rubric or not to rubric? The effects of self-assessment on self-regulation, performance and self-efficacy. *Assessment in Education, 21*(2), 133–148.

Pintrich, P. R., Smith, D. A., Garcia, T., & Mckeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. Ann Arbor: National Center for Research to Improve Postsecondary teaching and Learning, University of Michigan.

Primi, R., Wechsler, S. M., Nakano, T. C., Oakland, T., & Guzzo, R. S. L. (2014). Using item response theory methods with the Brazilian Temperament Scale for students. *Journal of Psychoeducational Assessment, 32*(7), 651–662.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement test*. Copenhagan: Danish Institute for Educational Research. Expanded ed. (1980). Chicago: The University of Chicago Press.

Richardson, J. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387–415.

Ryan, M. E. (2014). Reflexive writers: Rethinking writing development and assessment in schools. *Assessing Writing, 22,* 60–74.

Suh, H. N., Wang, K. T., & Arterberry, B. J. (2015). Development and initial validation of the self-directed learning inventory with Korean college students. *Journal of Psychoeducational Assessment, 33*(7), 687–697.

Swann, W. B., Jr., Pelham, B. W., & Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *Journal of Personality and Social Psychology, 57*(5), 782–791.

Tan, K. H. K. (2012). *Student self-assessment. Assessment, learning and empowerment*. Singapore: Research Publishing.

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., & 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: Erlbaum Associates.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest, version 2.0: Generalized item response modelling software*. Camberwell: Australian Council for Educational Research.

Yan, Z. (2016a). The self-assessment practices of Hong Kong secondary students: Findings with a new instrument. *Journal of Applied Measurement, 17*(3), 335–353.

Yan, Z. (2016b). Student self-assessment practices: The role of gender, year level, and goal orientation. *Assessment in Education*. https://doi.org/10.1080/0969594X.2016.1218324.

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262.