INTERNATIONAL JOURNAL OF Training and Development

International Journal of Training and Development 24:4
ISSN 1360-3736
doi: 10.1111/ijtd.12201

Psychometric properties of the Self-assessment Practice Scale for professional training contexts: evidence from confirmatory factor analysis and Rasch analysis

Zi Yan, Sonja Brubacher, David Boud and Martine Powell

Self-assessment is a fundamental skill for professionals because self-assessment can promote self-regulated learning and professional development. However, studies reporting the use of self-assessment instruments in the professional training context are scarce. This study aimed to re-evaluate the psychometric properties of the Self-assessment Practice Scale (SaPS), which was originally developed in the school context, and extend its use to the professional training context. A sample of 200 investigative interviewer trainees from Australia and North America were invited to complete the modified SaPS. After removing misfitting items, the confirmatory factor analysis results confirmed a first-order four-factor solution. The multidimensional Rasch analysis demonstrated that the resultant 16 items had satisfactory fit to the Rasch model. In general, the results supported the use of the 16-item modified SaPS as a valid measure for the sample in this study. The potential of using the SaPS in professional training contexts is discussed.

❒ Zi Yan, Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong, China. Email: zyan@eduhk.hk. Sonja Brubacher, Centre for Investigative Interviewing, Griffith Criminology Institute, Griffith University, Brisbane, Australia. Email: s.brubacher@griffith.edu.au. David Boud, Deakin University, Geelong, Australia. University of Technology, Sydney, Australia. Middlesex University, London, UK. Email: david.boud@deakin.edu.au. Martine Powell, Centre for Investigative Interviewing, Griffith Criminology Institute, Griffith University, Brisbane, Australia. Email: martine.powell@griffith.edu.au.

Introduction

Professionals, no matter what their area of expertise, should make continuing judgements about their own work if they are to be effective in their practice. These judgements, often termed self-assessments, require those professionals to be able to appraise what they do, how they do it and make changes accordingly (Tai *et al*., 2018). Self-assessment within educational contexts has attracted increasing research interest due to its promise in promoting self-regulated and lifelong learning (Andrade, 2010; Boud, 1995; Bourke, 2018; Yan & Brown, 2017). In higher and K-12 education, the positive correlation between self-assessment and learning performance has been well recognized (e.g. Andrade, 2019; Brown & Harris, 2013).

Self-assessment is also regarded as a crucial skill in vocational education and professional training because self-assessment skill can promote self-regulated learning and facilitate learners' professional development throughout their careers (Kersh *et al*., 2011; Panadero *et al*., 2018). However, professional learners who enrol in training at the behest of their organizations may not be motivated to change their behaviours (Powell *et al*., 2010) or may be motivated only by performance goal orientations. This orientation is linked to lower use of self-assessment (Yan, 2018a) and self-regulatory strategies compared to a mastery goal orientation (Pintrich & Schrauben, 1992). Such learners may be neither internally motivated nor able to identify weaknesses in their performance (Regehr & Eva, 2006). There are also some criticisms on the usefulness of self-assessment in professional training. For example, the self-assessment results were likely to be inaccurate (Hodges *et al*., 2001). Some studies in nursing and medical education, for instance, have suggested that the link between self-assessment and actual performance may not be prominent (e.g. Baxter & Norman, 2011; Regehr & Eva, 2006). These criticisms tend to come, however, from an oversimplified conception of self-assessment as 'grade guessing' (Boud & Falchikov, 1989) rather than a learning process. Such research focused on self-assessment as a substitute for summative assessment rather than as a skill to be developed over time (Boud *et al*., 2013).

Recently, a process perspective (that acknowledges the complexity of self-assessment) has gained increasing endorsement in both general education (Andrade & Du, 2007; Fastre *et al*., 2012; Yan & Brown, 2017) and professional training (Epstein *et al*., 2008; Sargeant *et al*., 2010). Yan (2020b) found self-assessment to be a fundamental skill for self-regulated learning which can occur at different self-regulated learning phases with different patterns and for different purposes. For example, at preparatory phase, self-assessment aims to identify personal and external resources and to set reasonable learning goals; at performance phase, self-assessment helps to monitor the learning process and to ensure that the learning strategies are appropriate for achieving goals; at appraisal phase, self-assessment can identify learners' strengths and weaknesses as well as the directions for future improvement. The understanding of such an intertwined relationship between self-assessment practices and self-regulated learning is particularly crucial in professional training because the success of professional training is heavily reliant on learners' willingness and competence for self-regulated learning.

One significant challenge in self-assessment literacy is a lack of common understanding of self-assessment processes and valid instruments specifically designed for assessing self-assessment practices (Panadero *et al*., 2016; Yan, 2018b). Compared with higher and K-12 education, self-assessment publications in professional training are limited and studies reporting the use of self-assessment instruments are extremely scarce (Panadero *et al*., 2018), especially those adopting a process perspective. The current study was conducted to determine whether a suitably adapted pre-existing instrument, normed on a school education population, would be a reliable tool in other contexts. Specifically, this paper aims to re-evaluate the psychometric properties of the Self-assessment Practice Scale (SaPS) for use with professional trainees and extend the use of the scale from the school context to the professional training context. Investigative interviewers were a desirable population for testing the generalizability of the SaPS because their work arguably requires ongoing assessment (in some format) for maintenance of skills (Lamb, 2016).

Defining self-assessment process

There is a general consensus that self-assessment is a complex process incorporating multiple steps, such as determining criteria and making judgements (Andrade *et al*., 2008; Boud, 1995). Some attempts have been made to define the self-assessment process (e.g. Fastre *et al*., 2012; McMillan & Hearn, 2008; Ross, 2006; Sargeant *et al*., 2010). However, those attempts have put limited attention to unpacking the inner process of self-assessment, i.e., what actions or steps learners conduct when they engage in self-assessment. As a response to this challenge, Yan and Brown (2017) recently 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 (p. 1248)'. Based on in-depth interviews with students from a Hong Kong teacher education institute, a cyclical process model was developed which identified three steps in the self-assessment process including determining assessment criteria, self-directed feedback seeking and self-reflection (see Figure 1).

Firstly, students decide the self-assessment criteria that will be used in subsequent steps. Secondly, they set out to collect feedback on their performance from external and/or internal sources. External feedback can be obtained from monitoring the learning process (e.g. doing extra work), and/or from inquiry with others (e.g. teachers and peers). Internal feedback comes from internal reactions (e.g. emotions and internal states) triggered by their own performance. Thirdly, students reflect on the quality of the learning process based on the feedback obtained in the second step with an aim to identify their own strengths and weaknesses. Following this process, students make an initial self-assessment judgement that is subjected to continuous calibration depending on new feedback and/or new assessment criteria.

The Self-assessment Practice Scale (SaPS)

Yan (2018b) developed the SaPS in the school context with a theory-driven approach. The cyclical process model proposed by Yan and Brown (2017) underpinned the scale development. This scale was designed to measure self-assessment actions so as to (1) better understand the inner processes of self-assessment; and (2) depict the characteristics of self-assessment actions. Scale data can be used to monitor learners' self-assessment process and inform instructional design that can promote productive self-assessment that optimizes learning.

The SaPS focuses on only two steps in the self-assessment process – self-directed feedback seeking and self-reflection. The first step – determining criteria – is not covered because, for most self-assessment scenarios, learners may engage in various activities pertaining to self-directed feedback seeking and self-reflection, but are likely to select only one criterion. The 20 items are grouped into four subscales corresponding to the four self-assessment actions: seeking external feedback through monitoring (SEFM, 5 items), seeking external feedback through inquiry (SEFI, 4 items), seeking internal feedback (SIF, 4 items) and self-reflection (SR, 7 items). A six-point Likert-type response scale ranging from 1 (Strongly disagree) to 6 (Strongly agree) was used. The scale was validated with a sample of 2,906 Hong Kong students aged between 9 and 14 years. Both the results of factor analysis and Rasch analysis supported the SaPS as a satisfactory measure for use with school students in Hong Kong. The results of confirmatory factor analysis (CFA) confirmed that items in the four subscales performed well as specified in Yan and Brown's (2017) theoretical model. By examining the item-level fit statistics, Rasch analysis further supported that the 20 items in the SaPS fit the Rasch model. In other words, all items are performing well as specified in the underlying theory. The Cronbach's alpha coefficients/Rasch reliabilities for the four subscales were all satisfactory: .85/.88 for SEFM; .84/.88 for SEFI; .79/.80 for SIF; and .90/.90 for SR. Apart from the original validation study, the use of the SaPS has been extended to graduate students and showed satisfactory psychometric properties (Yan, 2020b).

Figure 1: Theoretical model of the cyclical self-assessment process (Yan & Brown, 2017). SDFS = self-directed feedback seeking; SR = self-reflection.

Self-assessment practices of investigative interviewer trainees

The field of investigative interview training has a long history of research associated with effective interviewing practices. Yet, it has only recently begun to explore how best to train the requisite knowledge and skills so that they transfer to actual practice in the workplace (Lamb, 2016; Powell, 2008; Smets & Rispens, 2014). There is growing recognition that the traditional didactic model is not effective to train interviewing skills (e.g. Wimshurst & Ransley, 2007). Further, it is less suitable to adult learning styles than more andragogical approaches, which encourage learners to self-regulate, self-assess and take responsibility for their learning (Birzer, 2003; Knowles, 1980; Vodde, 2012). Because this shift in interviewer training is relatively recent, little is known about the efficacy of investigative interviewers in accurately self-assessing their skills. It is well known that without some kind of ongoing assessment (e.g. self, supervisory), effective interviewing skills have a tendency to decline to post-training levels (Lamb, 2016; Lamb *et al*., 2002; Smith *et al*., 2009).

One study using a sample of 19 law students found that consistent self-evaluation over a 10-week period, coupled with ongoing interviewing and peer evaluations, was effective in improving interviewing skill (Stolzenberg & Lyon, 2015). The contribution of self-assessment to performance change, however, cannot be disentangled from the other learning factors of feedback and ongoing practice. Further, the process by which learners self-assessed their interviews is unknown. There is a striking gap in the investigative interviewing literature concerning the process of self-assessment and its links to knowledge and skills transfer.

The current study

To re-evaluate whether the SaPS is a valid instrument in the professional training context, the investigation was guided by Messick's (1995) framework of validity. Messick suggested that validity should be examined from six aspects: i.e., content, substantive, structural, generalizability, external and consequential aspects of validity. In this study, the content aspect of validity was examined by a panel of experts including the study authors and two independent interviewer training professionals. The item content relevance and representativeness were checked, and necessary modifications were made to ensure that the instrument is appropriate for use with investigative interviewer trainees. The substantive aspect of validity focuses on the extent to which the items are reflective of the underlying theory. This was ensured by following a valid theoretical model, i.e., Yan and Brown's (2017) self-assessment process model, and further investigated by indicators from Rasch analysis, such as the step calibrations and the item fit statistics. CFA was employed to examine the structural aspect of validity. That is, whether items' empirical relations to one another are consistent with the theoretical specifications. Applying the SaPS, which originated from school contexts, to the context of professional training would provide evidence regarding the generalizability aspect of validity. Given the close relationship between self-assessment, feedback seeking and self-regulated learning, the external aspect was gauged by the correlations between the SaPS subscales and widely used measures of feedback orientation and self-regulated learning strategies. Due to the research design, the consequential aspect of validity could not be examined in this study.

Method

Participants

Participants were current and former investigative interviewer trainees from Australia and North America who took one of the blended learning programs (online and face-to-face components) developed by Centre for Investigative Interviewing, Griffith University. These programs were undertaken with practitioners to enhance their forensic interviewing skills with child witnesses. In such contexts, self-assessment skills are vital. Investigative interviewing is a complex skill because evidence

from such interviews has no evidentiary value if they are not conducted well and professional interviewers must demonstrate autonomy, judgement and responsibility within broad parameters (Powell *et al*., 2010). There were 402 trainees invited to participate via either an email invitation or a link at the conclusion of their training. Of those, 200 completed the SaPS survey fully (148 female and 52 male). Their professional backgrounds included policing (n = 133), child protection and social work (n = 21), forensic interviewing (n = 21), academics (n = 3), students (n = 2), managers (n = 5), psychologists (n = 12) and lawyers (n = 3). Their ages varied from 20-29 years (n = 25), 30-39 years (n = 83), 40-49 years (n = 67), 50-59 years (n = 22) and 60-69 years (n = 3). Participants' experience in their current profession ranged from 0 to 40 years (M = 13.07, SD = 8.14). The highest education level of participants included year 10 (n = 2), year 11 (n = 4), year 12 (n = 30), certificate or diploma (n = 57), bachelor degree (n = 53), honours (n = 6), graduate diploma (n = 11), master's degree (n = 32) and doctorate degree (n = 5). Participants read and signed informed consent statements. Ethics approval was sought and given by Griffith University.

Measures

The measures used in this study included the SaPS, the feedback seeking measures and the self-regulated learning measures. The latter two were used to check the external aspect of validity of the SaPS.

Self-assessment Practice Scale

The SaPS was used but some modifications were made to better reflect the learning experience of investigative interviewer trainees. One item in the SR subscale (i.e. 'I reflect on my weaknesses when I discuss study-related issues with my classmates') was removed because the learning took place in an online environment and many trainees were the sole enrollee at their organization. The wording for some items was also modified according to the target context (e.g. 'extra exercises' was changed to 'extra work'). A definition of self-assessment, revised from Yan and Brown's (2017) definition, was provided to the participants at the beginning of the survey aiming to formulate a common understanding of self-assessment among respondents:

> Self-assessment refers to processes in which learners actively identify yardsticks they can use to judge their work, seek feedback about their own performance and reflect on their work and their processes of learning to identify their own strengths and weaknesses.

Feedback seeking measures

The Feedback Orientation Scale (FOS) was developed by Linderbaum and Levy (2010) to assess an individual's overall receptivity to feedback. FOS has four subscales and two of them were used in this study. *Utility* subscale (5 items; α = .88) was used to assess one's belief that feedback is useful for obtaining desired outcomes. *Self-efficacy* subscale (5 items; α = .78) was used to assess one's perceived competence to interpret and use feedback appropriately. It was hypothesized that the three subscales relevant to feedback seeking (i.e. SEFM, SEFI and SIF) in SaPS would positively correlate with Utility and Self-efficacy subscales in FOS.

Self-regulated learning measures

Dunn *et al*. (2012) re-examined the two subscales in the Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich *et al*., 1991), which were specifically designed to assess self-regulation, i.e., metacognitive self-regulation and effort regulation subscale. They proposed two modified scales, namely the General Strategies for Learning (GSL) scale and the Clarification Strategies for Learning (CSL) scale, to assess academic self-regulation. The GSL (5 items, α = .74) refers to general self-regulation strategies. The CSL (3 items, α = .61) refers to strategies in clarifying confusion and misunderstandings identified in learning.

Data analysis

Two analytical approaches, i.e., CFA and Rasch analysis, were applied to provide comprehensive scrutiny of the psychometric properties of the modified SaPS. This combined analytical approach has been used in recent empirical studies (e.g. Hart *et al*., 2013; Primi *et al*., 2014; Testa *et al*., 2019; West *et al*., 2018; Yan, 2018b, 2020a) for the benefit of providing supplementary information about the psychometric properties of instruments.

CFA using AMOS 24.0 (Arbuckle, 2015) was applied to test the global fit between the empirical data and the hypothesized factor model. The model-data fit indices used included χ^2/df ratio, 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). An acceptable model fit is indicated by χ^2/df ratio of less than 3; values of .90 or higher for GFI and CFI; values of .08 or lower for RMSEA (McDonald & Ho, 2002); and values of .08 or lower for SRMR (Hu & Bentler, 1999).

Rasch analysis adopts a 'data fit the model' approach. That means the empirical data have to meet *a priori* requirements in order to achieve scientific measurement (Andrich, 2004; Bond & Fox, 2015). Rasch analysis checks the extent to which items in a scale reflect a unidimensional latent construct. Different from CFA that mainly examines global model-data fit, Rasch analysis can provide item-level fit statistics. The criteria used include response category functioning, Rasch reliability and item fit statistics (i.e. Infit MNSQ and Outfit MNSQ). Values of Infit and Outfit MNSQ between 0.75 and 1.33 are indicators of sufficient fit to the Rasch model (Wilson, 2005). As self-assessment practice was theoretically defined as a construct consisting of four different but inter-related actions (Yan & Brown, 2017), a multidimensional Rasch model (Adams *et al*., 1997) was considered more appropriate for the SaPS data than the unidimensional Rasch model. In a multidimensional Rasch model, all subscales are calibrated simultaneously and the correlations between the subscales are considered so that the measurement precision on each subscale is enhanced. ConQuest 2.0 (Wu *et al*., 2007) was employed for the analysis.

In addition, internal consistency was determined by Cronbach's alpha and Rasch reliability, which is an equivalent indicator of internal consistency in Rasch analysis. Although a Cronbach's alpha of 0.7 is frequently used as a criterion for good internal consistency, a value of .60 or greater was also regarded an acceptable level of internal consistency for a group-level assessment, especially for newly developed instruments (e.g. Dunn *et al*., 2012; Nunnally, 1967; Weiner *et al*., 2003).

Results

To examine the psychometric properties of the SaPS, the results from CFA are first outlined, followed by results regarding internal consistency of the SaPS, Rasch analysis and correlations between the SaPS subscales and relevant measures.

Confirmatory factor analysis

The value of Kaiser–Meyer–Olkin measure for the SaPS was 0.78, quite close to 0.8, and Bartlett's test of sphericity was $\chi^2(171) = 1080.303$, $p < 0.001$, indicating appropriateness of applying factor analysis on the data. CFA with maximum likelihood (ML) estimation was applied. In the initial validation study of the SaPS, Yan (2018b) reported that a higher-order factor model, which was in line with Yan and Brown's (2017) model, demonstrated a good model-data fit. In that model, SEFM and SEFI contributed to a second-order factor, namely seeking external feedback (SEF). SEF and SIF contributed to a higher-order factor, i.e., seeking feedback (SF), which, together with SR, constituted self-assessment (see Model 1 in Figure 2a). However, in a recent attempt to develop a short form of the SaPS, Yan (2020a) found that the loading of SEF on SF did not significantly deviate from unity, indicating that SEF might be redundant.

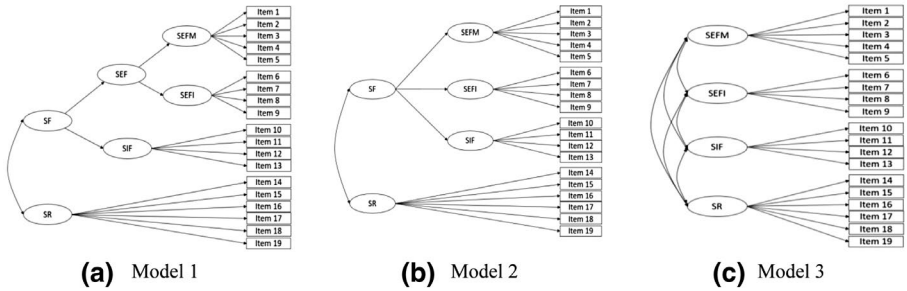

Figure 2: Three alternative models for the SaPS.

Thus, SEF was removed and a revised model (see Model 2 in Figure 2b) with SEFM, SEFI and SIF contributing to SF was preferred.

The CFA in this study started with Model 2. The results showed that the model-data fit was not satisfactory: $\chi^2/df = 2.734$; GFI = .806; CFI = .730; RMSEA = .093; and SRMR = .098. Examinations of the modification indices revealed that three items (Item #8 in SEFI; Item #13 in SIF; and Item #14 in SR) are the major sources of the misfit. Item #8 (I ask my colleagues to tell me how to improve my learning) appeared problematic possibly because in the sample a number of trainees were the sole person in their organization who was enrolled in the training. Although the question was meant to be general, they may have been thinking about the particular course, and so they did not have colleagues who they could ask for assistance with regard to learning this material. The malfunction of Item #13 (My intuition tells me if I am doing a good job or not) may have arisen because the course content explicitly challenged learners to reconsider their instinctual interviewing habits. Some learners may have interpreted this item in a general fashion, while others may have linked it to the specific course. The reason causing problems for Item #14 (I seek out the reasons for mistakes I made after getting back marked work) might be that, while trainees frequently received feedback during training, they might not have perceived it as traditional 'marked work' (e.g. like a school assignment). For example, they got verbal feedback during practice interviews and did quizzes which were automatically scored and the correct answer was provided along with a rationale.

After removing these three misfitting items, we re-ran the CFA and the model-data fit was much better (see Table 1). A more parsimonious first-order four-factor model (see Model 3 in Figure 2c) was also tested and the results showed that it had better model-data fit, as well as smaller Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC), than Model 2. Inspection of modification indices showed that, if the residuals of two pairs of items (Items #1 and #2; Items #4 and #17) were allowed to correlate, the fit arrived at a satisfactory level (see Model 3_modified in Table 1). The correlation between the residuals of the two pairs of items could be explained by the strongly related item contents: Items #1 (I check whether I have mastered course content by doing extra work) and #2 (I check whether I have fully understood course content by revisiting prior assignments) both involve comparing perceptions of internal learning states to external learning materials; Items #4 (I ask

Table 1: CFA goodness-of-fit indices for different models (three items removed)

	χ^2/df	GFI	CFI	RMSEA	SRMR	AIC	BIC
Model 2	1.981	.886	.863	.070	.079	270.085	388.824
Model 3	1.893	.895	.878	.067	.076	261.494	386.830
Model 3_Modified	1.691	.906	.908	.059	.072	242.288	374.221

myself questions in my head to check whether I have understood course content) and #17 (I think about whether the way I am studying is really helping me learn) are both about self-inquiry. Hence, the CFA results confirmed the first-order four-factor solution.

Internal consistency of the SaPS

Cronbach's alpha coefficients for SEFM, SEFI, SIF and SR were .71, .69, .54, and .77, respectively. Considering that each subscale has only three to five items, the results reflect acceptable internal consistency for SEFM, SEFI and SR. SIF appeared relatively weak and some further development might be necessary.

Rasch analysis

A multidimensional Rasch analysis was applied to the responses to the 19-item SaPS. The Rating Scale Model was used as all items shared the same set of response categories. The six-point response scale functioned well because the step calibrations (the measures of the transition points between adjacent categories) increased monotonically from −1.31, −0.59, −0.33, 0.56, to 1.67 logits. Table 2 shows that the correlations among SEFM, SEFI and SR were median to high, while the correlation between SIF and other latent traits was relatively low.

Inspection of item fit statistics identified one misfitting item (Item #8: Infit MNSQ = 1.68; Outfit MNSQ = 1.67). This result echoed the CFA results, indicating that Item #8 did not perform consistently with other items in the same subscale. To be in line with the CFA results, Items #13 and #14, together with Item #8, were removed although Items #13 and #14 did not show misfit in Rasch analysis. The Rasch analysis was re-conducted on the remaining 16 items. All items demonstrated satisfactory fit to the Rasch model. The item difficulty, standard error and item fit statistics are presented in Table 3. The Rasch reliabilities (i.e. EAP/PV reliabilities) for SEFM, SEFI, SIF and SR were .73, .75, .61 and .79. In line with Cronbach's alpha coefficients, SEFM, SEFI and SR demonstrated good Rasch reliabilities, while SIF had a marginally acceptable reliability.

The item-person map (Figure 3), also called Wright map, displays the hierarchy of measures with regard to item difficulties and person abilities. The distributions of person measures on each of the four subscales are presented in the four continua on the left side. Persons are placed from top to bottom with descending levels of self-assessment practice. The items are grouped into the four subscales and are placed on the right side, with the hardest items at the top and the easiest items at the bottom. It can be seen that the SaPS provided a fairly targeted measurement of respondents' self-assessment practices, although the range of item difficulty was smaller than the range of person ability.

Correlations with relevant measures

To examine the external aspect of validity of the SaPS, the Pearson correlations between Rasch-calibrated person measures on the four subscales of the SaPS and the person measures of the Utility and Self-efficacy subscales in FOS, CSL and GSL were calculated. Table 4 shows that, as expected, the person measures of the two subscales on

Table 2: Correlations between the four latent traits

	SEFM	SEFI	SIF	SR
SEFM	–			
SEFI	0.50	–		
SIF	0.17	0.48	–	
SR	0.60	0.60	0.39	–

Table 3: Item difficulty, standard error, item fit statistics for the 16-item SaPS

Scale/Item	Item measure*	SE	Infit MNSQ	Outfit MNSQ
Seeking external feedback through monitoring (*SEFM*)				
Item 1: I check whether I have mastered course content by doing extra work	0.736	0.051	0.94	0.97
Item 2: I check whether I have fully understood course content by revisiting prior assignments	0.211	0.051	1.09	1.1
Item 3: I keep track of my progress by recording my performance	0.319	0.051	0.95	0.93
Item 4: I ask myself questions in my head to check whether I have understood course content	−0.73	0.055	1.2	1.17
Item 5: I check my performance against answers in guides or on a website	−0.537	0.104	1.07	1.03
Seeking External Feedback Through Inquiry (*SEFI*)				
Item 6: I ask my teachers/trainers to give me feedback about my performance	−0.805	0.054	1.06	1.02
Item 7: I ask my family members to give me advice on my work	0.443	0.05	1.19	1.19
Item 9: I ask my fellow group members to evaluate my contributions to group work tasks	0.362	0.074	1	1.01
Seeking Internal Feedback (*SIF*)				
Item 10: My gut feelings tell me whether my work is good or bad	−0.175	0.048	1.13	1.1
Item 11: My emotions influence my evaluation of my learning performance	−0.203	0.048	0.97	0.97
Item 12: How my body feels tells me how well I am doing	0.377	0.068	1.04	1.07
Self-reflection (*SR*)				
Item 15: When others (e.g. teachers, family members, colleagues) make comments about my course work, I think about how much sense these make to me	0.146	0.058	1	1.09
Item 16: Any areas I am unsure of after finishing my work, I go over again	0.141	0.058	0.83	0.88
Item 17: I think about whether the way I am studying is really helping me learn	0.062	0.058	0.96	0.98

Table 3: (Continued)

Scale/Item	Item measure*	SE	Infit MNSQ	Outfit MNSQ
Item 18: When I do exercises, I look at what I got wrong or did poorly on to guide me as to what I should learn next	0.087	0.058	0.8	0.81
Item 19: I pay attention to my assessment results to identify what I can do better next time	−0.436	0.116	0.84	0.82

*All measures are in logits.

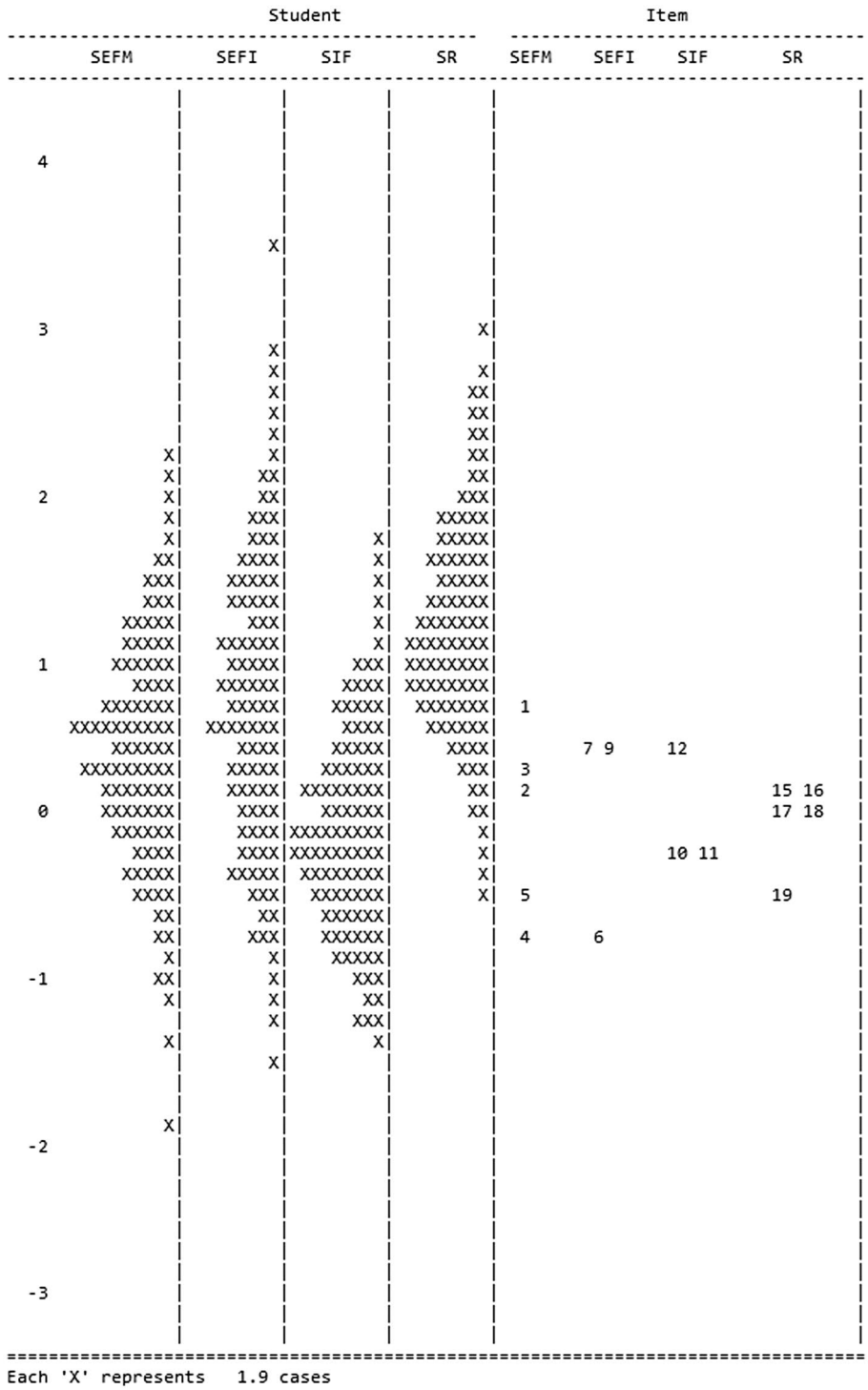

Figure 3: The wright map of the SaPS.

Table 4: Correlations between Person measures on the four SaPS subscales, two FOS subscales, CSL and GSL

	FOS_Utility	FOS_Self-efficacy	CSL	GSL
SEFM	.215**	.234**	.454**	.429**
SEFI	.441**	.251**	.332**	.288**
SIF	.127	.003	.064	.120
SR	–	–	.555**	.463**

feedback seeking (i.e. SEFM and SEFI) were significantly associated with the two subscales of FOS (correlation coefficients ranging from .215 to .441). The person measures of SEFM, SEFI and SR were significantly associated with CSL and GSL (correlation coefficients ranging from .288 to .555). However, the correlations between SIF and all other scales were low and non-significant.

Discussion and conclusion

This study set out to extend the use of the SaPS – an existing instrument previously normed on the school context – to the professional training context using a sample of investigative interviewer trainees. Three items were removed for substantive reasons, resulting in a scale with 16 items (SEFM – 5 items; SEFI – 3 items; SIF – 3 items; SR – 5 items). The data analysis results generally supported the use of the 16-item modified SaPS as a valid measure for the sample in this study.

The CFA results confirmed the first-order four-factor solution. The multidimensional Rasch analysis provided further support to the psychometric properties of the SaPS. All the 16 items in the resultant scale demonstrated satisfactory fit to the Rasch model, implying that items in each subscale are measuring a unidimensional latent trait. The six-point response scale functioned well, indicated by the ordered step calibrations. The SaPS items were well targeted at the respondents' self-assessment practices and covered a reasonable range along the latent trait scale. Both Cronbach's alpha coefficients and Rasch reliabilities showed that subscales SEFM, SEFI and SR had satisfactory reliabilities, while SIF appeared to be a less reliable subscale. The expected correlations with other relevant measures, such as feedback orientation and self-regulated learning strategies, provided evidence of the external aspect of validity of SEFM, SEFI and SR, but not for SIF. Furthermore, as the original SaPS was developed on the school context in a Confucian culture, its successful application on the professional training context in a Western (Australian and North American) culture gives support to its generalization across contexts and cultures.

Notwithstanding the promising findings concerning the generalizability of the SaPS, the SIF subscale warrants attention and probably further development in the professional training context. It showed the lowest reliability and correlations with other relevant measures compared to the other three subscales. While this subscale's performance was better in the school context than in the professional context of the current study, it was still the relatively weak one among all the four subscales (see Yan, 2018b). This subscale, in particular, may be strongly affected by the nature of the profession in which it is used. The SIF subscale assesses the degree to which learners evaluate the success of their learning by reflecting on their internal biophysiological states; this quality may be especially helpful in athletic and other physical professions, such as ballet and golf. This echoes Yan and Brown's (2017) findings that internal feedback was particularly salient for students majoring in performance-based disciplines, such as music education. In contrast, for tasks that rely on integrating feedback from external sources, such as other people, SIF items may be less applicable.

The consequential aspect of validity was not checked in this study. As self-assessment promises potentially adaptive learning outcomes, such as improved work quality

and self-regulated learning behaviour, future studies could fill in this gap by collecting evidence of intended and unintended consequences, associated with the interpretation and use of the SaPS measures (Messick, 1995).

The use of the SaPS in professional training

Self-regulated learning competency is crucial for professional trainees because they are required to identify ongoing workplace learning needs and respond to them (Munby *et al*., 2009). Self-assessment is regarded as a fundamental skill for self-regulated learning and plays an important role during the learning process (Andrade, 2010; Yan, 2020b; Yan *et al*., 2020). In a recent review study, Panadero *et al*. (2017) reported positive impacts of self-assessment interventions on learners' self-regulated learning strategies. Thus, the use of the SaPS has the potential to benefit professional training for two reasons. First, the data collected through the SaPS could assist in monitoring trainees' development of self-assessment skills and self-regulation competencies. Second, the characteristics of trainees' self-assessment practices revealed by the use of the SaPS could inform the design of professional training programs that encourage and facilitate ongoing assessment for maintenance of skills.

As an instrument assessing context-based behaviours, the SaPS provides a ready-for-use tool for researchers and professional trainers. But, more importantly, it serves as a useful framework for understanding and measuring self-assessment practices in different contexts. Because the focus of self-assessment might vary across courses and training contexts, learners' self-assessment practices might have different characteristics; for example, the sources from which learners seek feedback. Thus, some SaPS items were modified in this study to reflect the differences between the school context and the professional training context. For instance, 'extra exercises' and 'text books' were changed to 'extra work' and 'guides', respectively. However, the scale structure of the SaPS and the underlying theory (i.e. the self-assessment process model) remained invariant empirically. This suggests a flexible pattern for future applications of the SaPS in other contexts. That is, while a common SaPS scale structure underpinned by the self-assessment process model is necessary, it might be appropriate to make modifications to individual items to cater for diverse working environments of professional trainees. This pattern enables the use of the SaPS in a wide range of scenarios and, at the same time, ensures a common framework for understanding and interpretation of self-assessment practices across contexts.

The successful application of the SaPS in the current study also serves as a timely reminder that self-assessment in professional training should be regarded as a fundamental part of judging one's learning and developing confidence in it rather than as a prediction of marks others might give an exercise (Boud, 1995; Yan & Brown, 2017). The criticisms on self-assessment in terms of its inaccuracy (e.g. Hodges *et al*., 2001) and low correlation with actual performance (e.g. Baxter & Norman, 2011; Regehr & Eva, 2006) were largely built on an earlier misconception that self-assessment could be used as a substitute for the judgements of experts. From a pedagogical perspective, the benefits of self-assessment may come initially from the development of metacognition that results from reflective and active engagement in the learning process, rather than being considered from the start as 'veridical' or coinciding with reality (Yan & Brown, 2017). Thus, more emphasis should be put on whether learners have developed adaptive learning skills from self-assessment rather than always having their self-assessment results corroborated by the external standards (Andrade, 2010; Tan, 2012).

This study contributes to our understanding of investigative interviewer trainees' self-assessment practices and shows that the SaPS could be applied as a useful instrument in professional training contexts. Future research in this field could usefully explore whether and how the use of the SaPS leads to better training outcomes. On the one hand, such research could shed light on the consequential aspect of validity of the SaPS which was not examined here. On the other hand, the insights brought by the use of the SaPS have the potential to inform how long-term effects of professional training are maintained, which is an ever-lasting challenge associated with many professional training programs.

Acknowledgement

The first author was supported by a General Research Fund (GRF) (Project No: EDUHK 18600019) from the Research Grants Council of Hong Kong. The second, third, and fourth authors were supported by an Australian Research Council (ARC) Discovery Grant to DP180100715.

References

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

Andrade, H. L. (2010), 'Students as the definitive source of formative assessment: Academic self-assessment and the self-regulation of learning', in H. L. Andrade and G. J. Cizek (eds), *Handbook of formative assessment*, (New York, NY: Routledge), pp. 90–105.

Andrade, H. L. (2019), 'A critical review of research on student self-assessment', *Frontiers in Education*, **4**, 87.

Andrade, H. and Du, Y. (2007), 'Student responses to criteria-referenced self-assessment', *Assessment & Evaluation in Higher Education*, **32**, 159–81.

Andrade, H., Du, Y. and Wang, X. (2008), 'Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing', *Educational Measurement: Issues and Practice*, **27**, 3–13.

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

Arbuckle, J. L. (2015). *Amos* (*Version 24.0*) *[Computer Program]*. Chicago, IL: IBM SPSS.

Baxter, P. and Norman, G. (2011), 'Self-assessment or self deception? A lack of association between nursing students' self-assessment and performance', *Journal of Advanced Nursing*, **67**, 2406–13.

Birzer, M. L. (2003), 'The theory of andragogy applied to police training', *Policing: An International Journal of Police Strategies & Management*, **26**, 29–42.

Bond, T. G. and Fox, C. M. (2015), *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd edn (New York, NY: Routledge).

Boud, D. (1995), *Enhancing Learning Through Self-Assessment* (London, UK: Kogan Page).

Boud, D. and Falchikov, N. (1989), 'Quantitative studies of student self-assessment in higher-education: A critical analysis of findings', *Higher Education*, **18**, 5, 529–49.

Boud, D., Lawson, R. and Thompson, D. (2013), 'Does student engagement in self-assessment calibrate their judgement over time?', *Assessment and Evaluation in Higher Education*, **38**, 8, 941–56.

Bourke, R. (2018), 'Self-assessment to incite learning in higher education: Developing ontological awareness', *Assessment & Evaluation in Higher Education*, **43**, 5, 827–39.

Brown, G. T. L. and Harris, L. R. (2013), 'Student self-assessment', in J. H. McMillan (ed), *SAGE Handbook of Research on Classroom Assessment* (Thousand Oaks, CA: Sage), pp. 367–93.

Dunn, K., Lo, W.-J., Mulvenon, S. W. and Sutcliffe, R. (2012), 'Revisiting the Motivated Strategies for Learning Questionnaire: A theoretical and statistical reevaluation of the metacognitive self-regulation and effort regulation subscales', *Educational and Psychological Measurement*, **72**, 2, 312–31.

Epstein, R. M., Siegel, D. J. and Silberman, J. (2008), 'Self-monitoring in clinical practice: A challenge for medical educators', *Journal of Continuing Education in the Health Professions*, **28**, 5–13.

Fastre, G. M. J., van der Klink, M. R., Sluijsmans, D. and van Merrienboer, J. J. G. (2012), 'Drawing students' attention to relevant assessment criteria: Effects on self-assessment skills and performance', *Journal of Vocational Education & Training*, **64**, 2, 185–98.

Hart, C. O., Mueller, C. E., Royal, K. D. and Jones, M. H. (2013), 'Achievement goal validation among African American high school students: CFA and Rasch results', *Journal of Psychoeducational Assessment*, **31**, 3, 284–99.

Hodges, B., Regehr, G. and Martin, D. (2001), 'Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it', *Academic Medicine*, **76**, 87–9.

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

Kersh, N., Evans, K., Kontiainen, S. and Bailey, H. (2011), 'Use of conceptual models in self-evaluation of personal competences in learning and in planning for change', *International Journal of Training and Development*, **15**, 4, 290–305.

Knowles, M. S. (1980), *The modern practice of adult education: From pedagogy to andragogy*, 2nd edn (New York, NY: Cambridge Books).
Lamb, M. E. (2016), 'Difficulties translating research on forensic interview practices to practitioners: Finding water, leading horses, but can we get them to drink?', *American Psychologist*, **71**, 710–8.
Lamb, M. E., Sternberg, K. J., Orbach, Y., Esplin, P. W. and Mitchell, S. (2002), 'Is ongoing feedback necessary to maintain the quality of investigative interviews with allegedly abused children?', *Applied Developmental Science*, **6**, 35–41.
Linderbaum, B. A. and Levy, P. E. (2010), 'The development and validation of the Feedback Orientation Scale (FOS)', *Journal of Management*, **36**, 6, 1372–405.
McDonald, R. P. and Ho, M. R. (2002), 'Principles and practice in reporting structural equation analyses', *Psychological Methods*, **7**, 64–82.
McMillan, J. H. and Hearn, J. (2008), 'Student self-assessment: The key to stronger student motivation and higher achievement', *Educational Horizons*, **87**, 40–9.
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**, 741–9.
Munby, H., Hutchinson, N. L. and Chin, P. (2009), 'Workplace learning: Metacognitive strategies for learning in the knowledge economy', in R. Maclean and D. Wilson (eds), *International Handbook of Education for the Changing World of Work* (Netherlands: Springer), pp. 1763–75.
Nunnally, J. C. (1967), *Psychometric theory*, (New York, NY: McGraw-Hill).
Panadero, E., Brown, G. T. and Strijibos, J. W. (2016), 'The future of student self-assessment: A review of known unknowns and potential directions', *Educational Psychology Review*, **28**, 803–30.
Panadero, E., Garcia, D. and Fraile, J. (2018), 'Self-assessment for learning in vocational education and training', in S. McGrath, M. Mulder, J. Papier and R. Suart (eds), *Handbook of Vocational Education and Training: Developments in the Changing World of Work*, (Cham: Springer International Publishing), pp. 1–12.
Panadero, E., Jonsson, A. and Botella, J. (2017), 'Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses', *Educational Research Review*, **22**, 74–98.
Pintrich, P. R. and Schrauben, B. (1992), 'Students' motivational beliefs and their cognitive engagement in classroom academic tasks', *Student Perceptions in the Classroom*, **7**, 149–83.
Pintrich, P. R., Smith, D. F., Garcia, T. and McKeachie, W. J. (1991), *A Manual for the Use of the Motivated Strategies for Learning Questionnaire (MSLQ)*, (Ann Arbor, MI: University of Michigan).
Powell, M. B. (2008), 'Designing effective training programs for investigative interviewers of children', *Current Issues in Criminal Justice*, **20**, 189–208.
Powell, M. B., Wright, R. and Clark, S. (2010), 'Improving the competency of police officers in conducting investigative interviews with children', *Police Practice and Research*, **11**, 211–26.
Primi, R., Wechsler, S. M., Nakano, T. C., Oakland, T. and Guzzo, R. S. L. (2014), 'Using item response theory methods with the Brazilian Temperament Scale for students', *Journal of Psychoeducational Assessment*, **32**, 7, 651–62.
Regehr, G. and Eva, K. (2006), 'Self-assessment, self-direction, and the self-regulating professional', *Clinical Orthopaedics and Related Research*, **449**, 34–8.
Ross, J. A. (2006), 'The reliability, validity, and utility of self-assessment', *Practical Assessment*, **11**, 10, 1–13.
Sargeant, J., Armson, H., Chesluk, B., Dornan, T., Eva, K., Holmboe, E., Lockyer, J., Loney, E., Mann, K. and van der Vleuten, C. (2010), 'The processes and dimensions of informed self-assessment: A conceptual model', *Academic Medicine*, **85**, 7, 1212–20.
Smets, L. and Rispens, I. (2014), 'Investigative interviewing and training: The investigative interviewer apprentice', in R. Bull (ed), *Investigative Interviewing*, (New York, NY: Springer), pp. 147–65.
Smith, R., Powell, M. B. and Lum, J. (2009), 'The relationship between job status, interviewing experience, gender, and police officers' adherence to open-ended questions', *Legal and Criminological Psychology*, **14**, 51–63.
Stolzenberg, S. N. and Lyon, T. D. (2015), 'Repeated self-and peer-review leads to continuous improvement in child interviewing performance', *Journal of Forensic Social Work*, **5**, 20–8.
Tai, J., Ajjawi, R., Boud, D., Dawson, P. and Panadero, E. (2018), 'Developing evaluative judgement: enabling students to make decisions about the quality of work', *Higher Education*, **76**, 467–81.
Tan, K. H. K. (2012), *Student Self-assessment: Assessment, Learning and Empowerment* (Singapore: Research Publishing).

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., di Uccio, U. S., Trani, F. and Zappia, A. (2019), 'Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis', *International Journal of Science Education*, **41**, 3, 388–417.
Vodde, R. F. (2012), 'Changing paradigms in police training: Transitioning from a traditional to an andragogical model', in M. R. Haberfeld, C. A. Clarke and D. L. Sheehan (eds), *Police Organization and Training: Innovations in Research and Practice*, (Cham, Switzerland: Springer), pp. 27–44.
Weiner, I. B., D. K. Freedheim, J. R. Graham and J. A. Naglieri (eds). (2003), *Handbook of Psychology: Assessment Psychology* (Hoboken, NJ: Wiley).
West, C., Baker, A., Ehrich, J. F., Woodcock, S., Bokosmaty, S., Howard, S. J. and Eady, M. J. (2018), 'Teacher Disposition Scale (TDS): Construction and psychometric validation', *Journal of Further and Higher Education*, **44**, 2, 185–200. https://doi.org/10.1080/03098 77X.2018.1527022
Wilson, M. (2005), *Constructing measures: An item response modeling approach* (Mahwah, NJ: Erlbaum Associates).
Wimshurst, K. and Ransley, J. (2007), 'Police education and the university sector: contrasting models from the Australian experience', *Journal of Criminal Justice Education*, **18**, 106–22.
Wu, M. L., Adams, R. J., Wilson, M. R. and Haldane, S. A. (2007), *ACER ConQuest, version 2.0: Generalized Item Response Modelling Software* (Camberwell: Australian Council for Educational Research).
Yan, Z. (2018a), 'Student self-assessment practices: the role of gender, school level and goal orientation', *Assessment in Education: Principles, Policy & Practice*, **25**, 2, 183–99.
Yan, Z. (2018b), 'The self-assessment practice scale (SaPS) for students: Development and psychometric studies', *The Asia-Pacific Education Researcher*, **27**, 2, 123–35.
Yan, Z. (2020a), 'Developing a short form of the self-assessment practices scale: Psychometric evidence', *Frontiers in Education*, **4**, 153. https://doi.org/10.3389/feduc.2019.00153.
Yan, Z. (2020b), 'Self-assessment in the process of self-regulated learning and its relationship with academic achievement', *Assessment & Evaluation in Higher Education*, **45**, 2, 224–38. https://doi.org/10.1080/02602938.2019.1629390.
Yan, Z. and 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, 1247–62.
Yan, Z., Brown, G. T. L., Lee, C. K. J. and Qiu, X. L. (2020), 'Student self-assessment: Why do they do it?', *Educational Psychology*, **40**, 4, 509–32. https://doi.org/10.1080/01443410.2019.1672038