Applications of Rasch Measurement in Research and Practice

Zi Yan

zyan@eduhk.hk
Department of Curriculum and Instruction,
The Education University of Hong Kong
(Hong Kong, China)

ABSTRACT: *A growing number of studies utilising the Rasch model have been published, spanning various disciplines. This paper delves into five key domains where Rasch measurement is applied, namely: developing new instruments, creating short-form instruments, developing vertical scales, combining Rasch analysis and path analysis, and applications to classroom testing. Within each area, this paper presents research studies as illustrations of how Rasch measurement can effectively address measurement issues and advance practices. These studies provide concrete examples for reflection, highlighting typical procedures and potential pitfalls to avoid when employing the Rasch model for different purposes in diverse contexts.*

KEYWORDS: **Rasch Measurement, Rasch model in Research, Rasch model in practice.**

Received 12/05/2023 Revised manuscript received 30/06/2023 Published 30/12/2023.

1. Introduction

There is an increasing number of studies published using the Rasch model across a remarkable breadth of disciplines (Bondet al., 2020; Aryadoust et al., 2019) provided a bibliographic review of studies that applied Rasch measurement. Their review identified over 5,000 publications over the period from 1972 to 2019 that covered a wide range of areas, including "Rehabilitation (13.09%), education and educational research (11.97%), health care sciences services (9.84%), psychology/ mathematics (8.80%), and health policy services (6.88%) constitute the top five disciplines with the highest application of the Rasch model." (2019, p. 3) This article focuses on five areas of application of Rasch measurement, including developing new instruments, creating short-form instruments, developing vertical scales, combining Rasch analysis and path analysis, and applications to classroom testing. In each area, I report my own studies to demonstrate how the application of Rasch measurement helps solve measurement problems and advance practices. It should be noted that the selected areas are not a comprehensive list of Rasch applications. Furthermore, these included studies are not necessarily exemplars, but serve as concrete samples for reflection in terms of the typical procedures, and the pitfalls we need to avoid when applying the Rasch model for different purposes in various contexts.

2. Developing new instruments

Developing an instrument from the Rasch perspective is probably the most intuitive approach for applying the Rasch model for most practitioners. The procedure for applying the Rasch model for instrument validation is quite straightforward and standardised. Many problems encountered by researchers in the process of instrument validation come from flaws in their theoretical framework or data collection procedure, but not Rasch analysis itself. In other words, the instrument development and validation process has to start with a solid theoretical framework, followed by rigorous data collection and then meaningful Rasch analysis. The reason is very simple: you cannot create diamonds from rubbish. If the data are substandard, even Rasch analysis can't help. This section presents an example to illustrate how to use Rasch analysis to examine the psychometric properties of a theory-driven instrument. The Self-assessment Practice Scale (SaPS-20) was developed according to the model of the Cyclical Self-assessment Process (Figure 1, Yan & Brown, 2017, p. 1255), which could explicitly outline the concrete and sequential actions within the self-assessment process.

When engaging in self-assessment, students first determine the assessment criteria. They then seek feedback regarding their learning from various sources, which could be classified into two major categories, i.e., external and internal sources. Internal feedback refers to internally generated reactions to their own performance, such as emotions, physical sensations, and internal states. In contrast, 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). 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. As students tend to use only one assessment criterion in the same self-assessment process, it is less meaningful to develop a scale for determining performance criteria from the perspective of scale development. Thus, the SaPS focuses on the remaining key actions, i.e., self-directed feedback seeking and self-reflection. The item development followed a standard procedure, including item crafting, expert review, item revision, and pilot study.

One more thing that needs to be kept in mind when validating an instrument is to follow a solid validity framework to check the instrument's quality. In many cases, researchers examine the validity in a very *ad hoc* approach. But validation needs to be done in a systematic way because validity is a multi-faceted concept and each of the facets needs to be carefully investigated. Messick's renowned framework (1995) was utilised for the development and validation of the SaPS. Hence, the six aspects of validity in Messick's framework (1995) were examined against the evidence provided by Rasch analysis as far as possible.

The original SaPS-20 has four scales with 22 six-point Likert-type items. This version was administered to 2,906 Hong Kong primary and secondary school students. As self-assessment practice consists of different but interrelated actions, a multi-dimensional Rasch model (Adams et al., 1997) was applied to the SaPS-20 data rather than the unidimensional Rasch model. ConQuest 2.0 (Wu et al., 2007) was employed for the Rasch analysis. The initial Rasch analysis identified two misfitting items. These two items were removed for substantive reasons and the Rasch analysis was re-run. We have checked various indicators provided by the Rasch analysis, including step calibrations, item fit statistics, differential item functioning (DIF) across gender and year levels, Rasch reliability, and item-person alignment.

Most of the indicators of instrument quality were good. The six-point response scale functioned well because the step calibrations

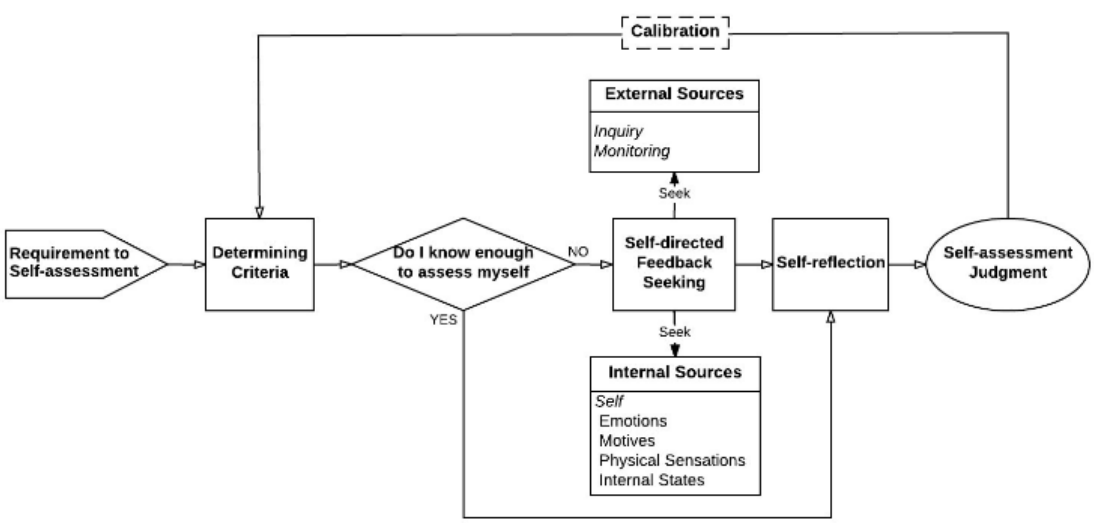


Figure 1. The cyclical self-assessment process (Yan & Brown, 2017, p. 1255)

Table 1. Item statistics in Rasch analysis (Yan, 2018, p. 131)

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

Figure 2. Wright map for the 20-item SaPS (Yan, 2018, p. 132)

(the measures of the transition points between adjacent categories) increased monotonically from -1.28, -1.14, -0.78, 0.87, to 2.34 logits. All items demonstrated sufficient fit to the Rasch model. DIF analyses showed that some minor DIF appeared on one item. The item difficulty, standard error, item fit statistics, and DIF results are presented in Table 1. The Rasch reliabilities (i.e. EAP/PV reliabilities) for each subscale, Seeking External Feedback Monitoring

(SEFM), Seeking External Feedback Inquiry (SEFI), Seeking Internal Feedback (SIF) and Self-Reflection (SR) were .85, .84, .79 and .90, respectively.

The Wright map (Figure 2), or item-person map, displays the hierarchy of measures with regard to item difficulties and person abilities. Although the range of the item difficulty is much smaller than the range of students' ability, these items, together with the response categories, provided a fairly targeted measurement of respondents' self-assessment practices.

3. Creating Short-Form Instruments

Contemporary social science research usually combines several instruments, instead of merely using one instrument in a study, to collect data on different variables, especially when an SEM model is an expected outcome. However, combining instruments tends to result in a lengthy questionnaire, which might increase respondents' workload and reduce the response rate. Thus, it is often beneficial to have short-form instruments with fewer items but still have satisfactory psychometric properties. To demonstrate the process of short-form instrument development, this section will utilise the aforementioned SaPS-20 as an example, using the same data set to develop a short form of it (hereafter SaPS-SF).

The first step was to select items from the original SaPS-20. The selected items should: (1) represent important content in terms of self-assessment practice; (2) have the largest structure coefficients within each of the four subscales; (3) have a good fit to the Rasch model; and (4) cover as wide as possible a difficulty range along the latent trait scale. After selecting items according to these four criteria, 12 items with 3 items in each of the 4 subscales were retained. Similar to the development of SaPS-20, these retained items were then subjected to confirmatory factor analysis (CFA) and Rasch analysis to examine their psychometric properties. To examine the invariance of estimates across the original scale (SaPS-20) and the short form (SaPS-SF), person invariance plots for each subscale were generated.

The items in SaPS-SF were first subject to CFA with maximum likelihood estimation. Surprisingly, SaPS-SF demonstrated a slightly better statistical fit than did the SaPS-20. Then, student responses to the items in SaPS-SF were subject to a multi-dimensional Rasch analysis. The step calibrations performed very well, increasing monotonically from -1.47, -1.27, -.76, .93 to 2.57 logits. The psychometric indicators for the SaPS-SF from CFA and Rasch analysis can be found in Table 2.

Table 2. Psychometric indicators for the SaPS-SF from CFA and Rasch analysis (Yan, 2020, p. 6)

Scale/Item	Item Measure*	SE	Infit MNSQ	Outfit MNSQ
Seeking External Feedback Monitoring (SEFM)				
Item 2. I check whether I have fully understood the course content by doing past exam papers.	-0.13	0.02	0.83	0.82
Item 1. I check whether I have mastered the course content by doing extra exercises.	-0.05	0.02	0.90	0.89
Item 3. I keep track of my progress by recording my performance.	0.18	0.03	1.01	1.01
Seeking External Feedback Inquiry (SEFI)				
Item 9. I ask my fellow group members to evaluate my contributions to group work tasks.	-0.03	0.03	0.97	0.96
Item 8. I ask my friends to tell me how to improve my learning.	-0.11	0.02	1.15	1.13
Item 6. I ask my teachers to give me feedback about my performance.	0.14	0.02	1.11	1.12
Seeking Internal Feedback (SIF)				
Item 12. How my body feels tells me how well I am doing.	0.08	0.02	1.00	1.01
Item 13. My intuition tells me if I am doing a good job or not.	-0.05	0.03	1.13	1.10
Item 10. My gut feelings tell me whether my work is good or bad.	-0.03	0.02	1.20	1.15
Self-Reflection (SR)				
Item 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.	0.15	0.02	0.80	0.83
Item 17. As I study, I think about whether the way I am studying is really helping me learn.	0.22	0.02	0.90	0.92
Item 19. I pay attention to my assessment results in order to identify what I can do better next time.	-0.37	0.03	1.12	1.07

Note. *All Rasch measures are in logits.

As shown in Table 3, both the conventional reliability (i.e., Cronbach's Alpha) and the Rasch reliability (i.e., EAP/PV reliabilities) remained satisfactory even though the number of items dropped substantially in each subscale. The separation reliabilities (i.e., person separation index) for SEFM, SEFI and SIF in both instruments were quite similar. Considering the considerable decrease in the item number, the scale reliability drop was insignificant.

The Wright map (Figure 3) shows that the range of the item difficulties seemed reasonable compared with the distribution of the student ability.

Further evidence indicating the similar functioning of the SaPS-SF compared with SaPS-20 could be found in the person invariance plots (Figure 4). The person measures of SaPS-SF were plotted against those generated from SaPS-20, with measures from SaPS-SF on the y-axis and measures from SaPS-20 on the x-axis. It's not difficult to discern that the person measures largely remain invariant across the two instruments since the person measures on all four subscales were mostly within the 95% control lines.

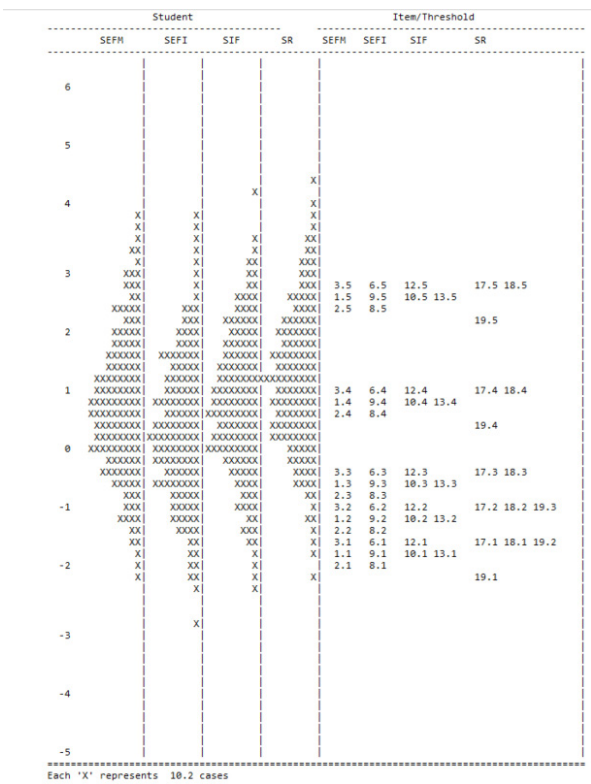


Figure 3. The Wright map of the SaPS-SF (Yan, 2020, p. 7)

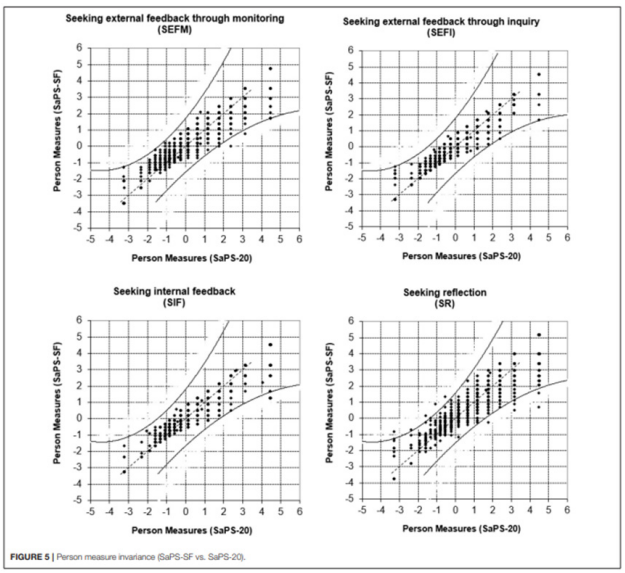


Figure 4. Person measure invariance (SaPS-SF vs. SaPS-20) (Yan, 2020, p. 8)

4. Developing Vertical Scales

Vertical scaling is a promising tool for tracking students' academic growth, which has drawn considerable attention among researchers and practitioners. Nevertheless, it is quite a challenging task to accomplish. There is no consensus regarding which approach is the optimal one to address the complexity of vertical scale construction. In this section, I will introduce the work of Yan et at., (2013), where they proposed a concurrent-separate approach based

Table 3. Comparison of reliabilities of the SaPS-20 and SaPS-SF (Yan, 2020, p.8)

	SaPS-20				SaPS-SF			
	SEFM	**SEFI**	**SIF**	**SR**	**SEFM**	**SEFI**	**SIF**	**SR**
Number of items	5	4	4	7	3	3	3	3
Cronbach's α	0.85	0.84	0.79	0.90	0.82	0.80	0.76	0.82
EAP/PV Rasch reliability	0.88	0.88	0.80	0.90	0.85	0.86	0.79	0.84
Person separation index	2.71	2.71	2.00	3.00	2.38	2.48	1.94	2.29

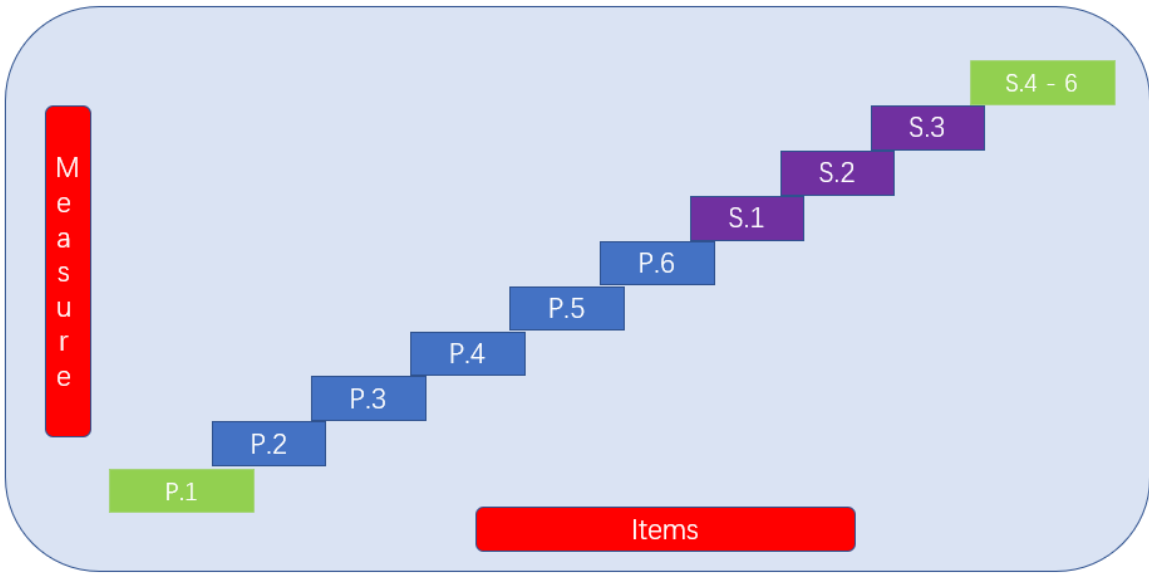


Figure 5. Assessment design for the scale (Yan et al., 2013, p. 193)

on Rasch analysis in developing a vertical scale, the Mathematics Competency Vertical Scale (MCVS) (Figure 5), to measure the development of Hong Kong students' competencies in mathematics.

The development of MCVS was based on a large sample of 9,531 students across primary 2 to secondary 3 (Table 4). For each grade, two assessment booklets (i.e., one for students who had just completed their first semester and the other for those who had just completed their second semester) were designed. Most items in the booklet were different, but for the adjacent two booklets, about 15% of their items were identical. For example, assuming a set of 50 items for primary 2 students who have just completed their second semester, approximately 7 or 8 items might be the same as those for primary 3 students who have just completed their first semester. These identical items are called common items or linking items, and this is called common item design. Through the common item design and subsequent Rasch analysis, the competencies of students from different grade levels become comparable within a single measurement framework.

Table 4. The item and participant distribution of booklets (Yan et al., 2013, p. 194)

Booklet	Number of item	Number of participant
P2_1	47	659
P2_2	42	650
P3_1	31	515
P3_2	35	514
P4_1	36	380
P4_2	36	382
P5_1	36	862
P5_2	36	756
P6_1	35	495
P6_2	36	542
S1_1	29	382
S1_2	35	227
S2_1	31	1405
S2_2	34	1393
S3_1	31	192
S3_2	32	177
Total	562	9531

A two-step analysis method was applied to the data. Step 1 aims to identify suitably qualified linking items for each level with separate analyses. This is very important because only through linking items of sufficient quality, can a stable framework within which student performance could be compared directly be built. In Step 2, data were stacked together according to the identified linking items and a Rasch analysis was conducted on all data at once to get item difficulty measures.

Within Step 1, separate Rasch analyses were conducted for each booklet. Underfitting persons whose OUTFIT or INFIT MNSQ were larger than 2.0 were identified and removed since they had a substantially negative impact on Rasch analysis results (Linacre, 2011). Then, analysis was run again to check whether the quality of

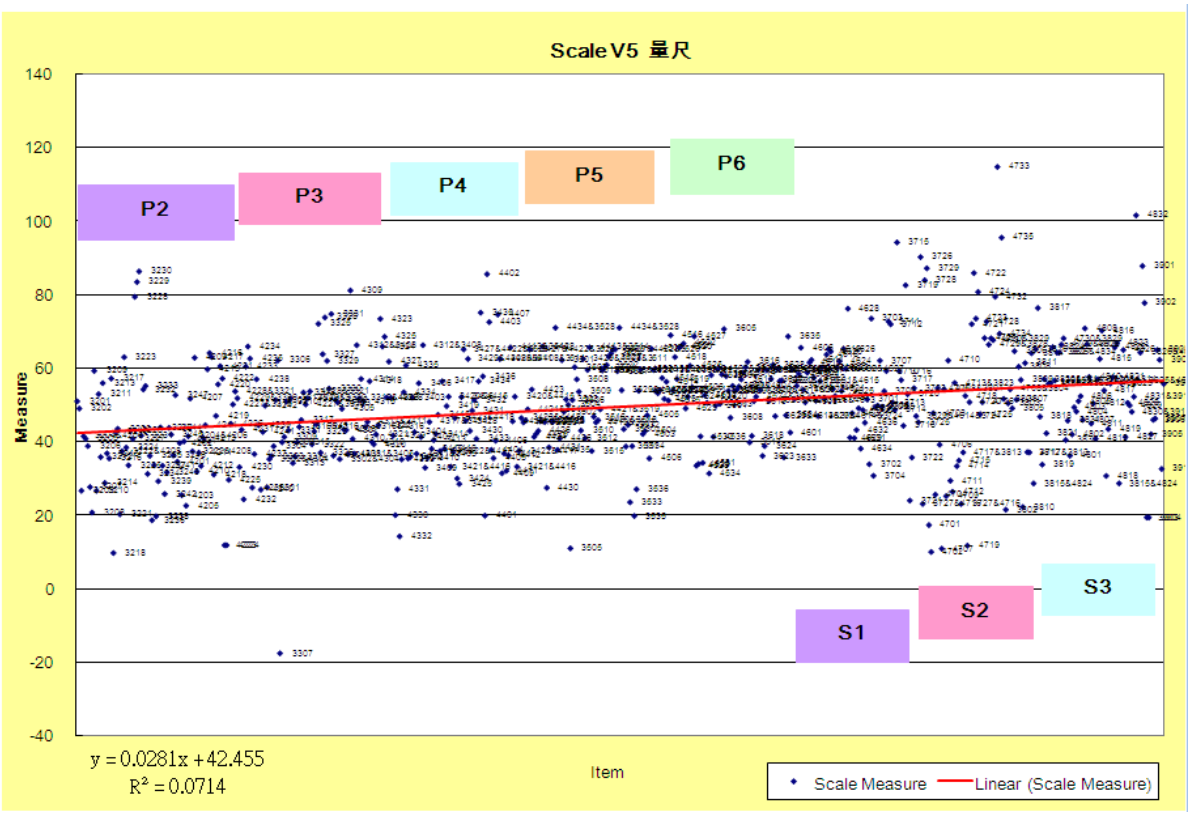


Figure 6. A vertical scale for mathematics (Yan et al., 2013, p.197)

the linking items fit the Rasch model. Because of the separate analyses, each linking item then had two separate item difficulty estimates. So, it was necessary to ensure these two separate item measures were comparable across the two adjacent grades. Linking items that failed either of the following two criteria were marked as disqualified and treated as different items in the following step: (1) demonstrating good fit to the Rasch model (i.e., the OUTFIT or INFIT MNSQ of the item was between .5 and 1.5); (2) remaining invariant across adjacent grades (i.e., the standardised difference of the item difficulties for adjacent grades was smaller than 2.0, and the actual difference of the item difficulties was smaller than .5 logits).

As a result, 37 qualified linking items were identified. In Step 2, all the data were stacked together by putting the student responses to each qualified linking item in the same column. Disqualified linking items were treated as different items even though they were the same in the content, and student responses to these disqualified items were put in different columns. By so doing, all the data from primary 2 to secondary 3 have been put together with the qualified linking items acting as anchors across grades. And then again, two rounds of analyses (i.e., identifying and removing the underfitting persons, and re-running Rasch analysis) were conducted. In this way, all the items were calibrated simultaneously onto a single latent trait scale.

The result of the above procedure is shown in Figure 6. Each dot in Figure 6 stands for a single item. A total of 510 unique items were retained in the final version of the MCVS. The items are grouped by their grades and placed along the *x*-axis from the left to the right. The *y*-axis represents item difficulty. The red solid line is a regression line that indicates that the item difficulty could be predicted, to some extent, by the grade where the item is placed. The item difficulty increased gradually in this vertical scale. Nevertheless, the remaining spread of many of the items and the regression line indicate that, in spite of the care taken, successful vertical scaling remains an arduous task.

5. Combining the Rasch and Path Analysis

For many social science researchers, structural equation modelling (SEM) appears as their default choice for data analysis. Rasch analysis does not replace conventional statistical analyses, such as SEM, but it can help solve fundamental measurement issues, such as transferring ordinal data into interval data, which can be subsequently analysed by SEM. In this sense, a "Rasch + Path analysis" approach warrants more attention and discussion.

As explained in the latest edition of Applying the Rasch Model (hereafter ARM4) (Bond et al., 2020), this solution divides analysis into two steps, first, the Rasch analysis and then, the path analysis. Rasch analysis at first can help ensure the quality of data for each variable that will be used in the subsequent path analysis to examine the relationship between the different latent traits.

One problem with this approach is that it does not take into account the standard errors of measurement associated with Rasch measures. As we are aware, Rasch measures are estimations, which are subject to errors. Hence, subsequent analyses should to take these errors into account. However, there is no SEM software, at least for now, that allows the user to input the error terms into analysis along with the Rasch measures. So, in this case, if we simply follow the sequence of Rasch analysis + Path analysis, we can input only the Rasch measures for each variable in order to do the Path analysis and no consideration can be given to the errors. Although this straightforward approach was used in some published papers, this impasse remains unsolved.

A recently published paper (Yan et al., 2020) could serve as an example of attempts trying to solve the problem, in which, five sets of plausible values were first generated through *ConQuest.* These sets of values were then input into the SEM software *Mplus*, which then ran the path analysis five times and automatically averaged to produce the final results. Of course, this approach is not ideal for solving the problem since yet as the error terms were not input directly. It is a compromise, but at least it took into account the error influence to some extent by having five sets of plausible values rather than relying on only one single set of estimates.

More promisingly, Moritz Heene, a co-author of ARM4, has innovatively proposed another solution which might offer a better option. He suggested incorporating the Rasch model in the SEM framework (see Figure 7). A typical SEM framework consists of two models: the measurement model and the structural model. Conventionally, CFA is used in the measurement model, and path Analysis is used in the structural model. Heene's proposal is to specify a Rasch analysis to replace CFA in the SEM framework.

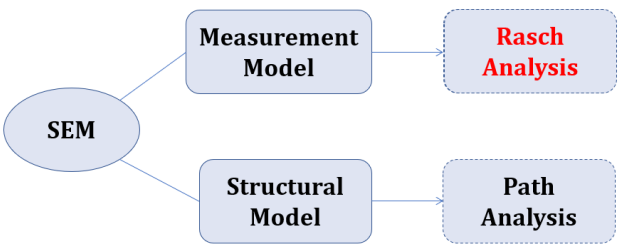
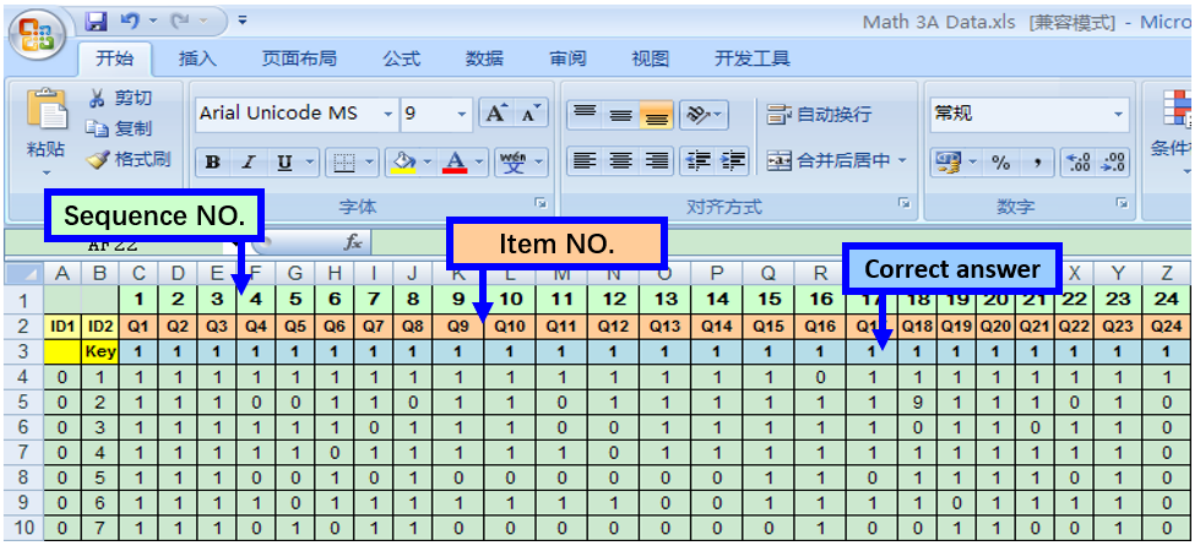


Figure 7. Incorporate the Rasch model in the SEM framework

Theoretically, a specified Rasch model for the measurement model with simultaneous path analysis can solve the measurement error issue because the path analysis is based on the Rasch measures as the latent trait, not the observed scores. But this approach brings a practical problem. Specifying a Rasch model in the measurement model requires that the corresponding identical slope assumption (i.e., identical item discrimination) should be implemented. However, when that identical slope specification is built into the measurement model, the requirement is so rigid that the vast majority of practical datasets could never fulfil it. This will result in unacceptable fit statistics and, therefore, the model is then very likely to be rejected. So, we are still in a dilemma. Given the wide adoption of Rasch analysis and SEM in social science research, more work is necessary in this line of inquiry.

6. Applications to Classroom Testing

While the previous four applications are research-oriented, this section will explore a more practice-oriented area, i.e., how Rasch analysis could be used in classroom testing. Applying Rasch analysis to classroom testing



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	ID1	ID2	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24
3		Key	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1
5	0	2	1	1	1	0	0	1	1	0	1	1	0	1	1	1	1	1	1	9	1	1	1	0	1	0
6	0	3	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	1	1	0	1	1	0	1	1	0
7	0	4	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0
8	0	5	1	1	1	0	0	1	0	1	0	0	0	0	0	0	1	1	0	1	1	1	1	0	1	0
9	0	6	1	1	1	1	0	1	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0
10	0	7	1	1	1	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0

Figure 8. An example of using Excel to teach Rasch analysis

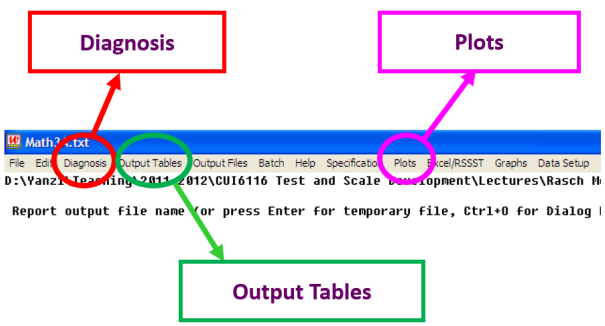


Figure 9. Three key functions of Winsteps

can inform learning and teaching by providing detailed, individualised information about student performance and item quality. These are areas with great potential and importance because what Rasch analysis can offer the well-informed classroom teacher is well aligned with the assessment for learning movement in education.

The key players in the process of Rasch application in the classroom must be the teachers. I have run a professional development programme entitled "Assessment Literacy and Effective Use of Assessment Data" for in-service teachers in Hong Kong for six years. The aim of this programme is to train teachers to retrieve valuable information from their own assessment data through Rasch analysis to serve the assessment for learning purposes.

Teachers start the data analysis from an Excel worksheet (Figure 8) with data (including information such as sequence numbers, item numbers and correct answers). With necessary instruction, teachers can analyse the data in either *Winsteps* (http://www.winsteps.com) or *Ministep* (http://www.winsteps.com/ministep.htm) (a free trial version of Winsteps).

Although *Winsteps/Ministep* offers a plethora of functions to generate various forms of outputs, I usually focus on a couple of indicators under three functions: Output Tables, Diagnosis, and Plots (Figure 9), as too much information could easily overwhelm teachers and eventually hinder their intention to use this method. In the subsequent sections, I will demonstrate how two indicators (i.e., the Variable maps and the PKMAPs) are introduced to teachers.

6.1. The variable map

The variable map can address the targeting issue in classrooms. The targeting issue here refers specifically to the alignment between student performance and item difficulty. Examining the alignment can help identify each student's Zone of Proximal Development (ZPD) and, therefore, inform teaching modifications according to each student's learning needs. For example, in Figure 10, the ZPD of students No. 8, 20, 28, 10, 18, and 19 is likely falling in the area, in terms of learning content and difficulty level, represented by Questions No. 10, 12, 22, 17, 8 and 9. Thus, when helping these students (e.g., developing follow-up worksheets), exercises or practices

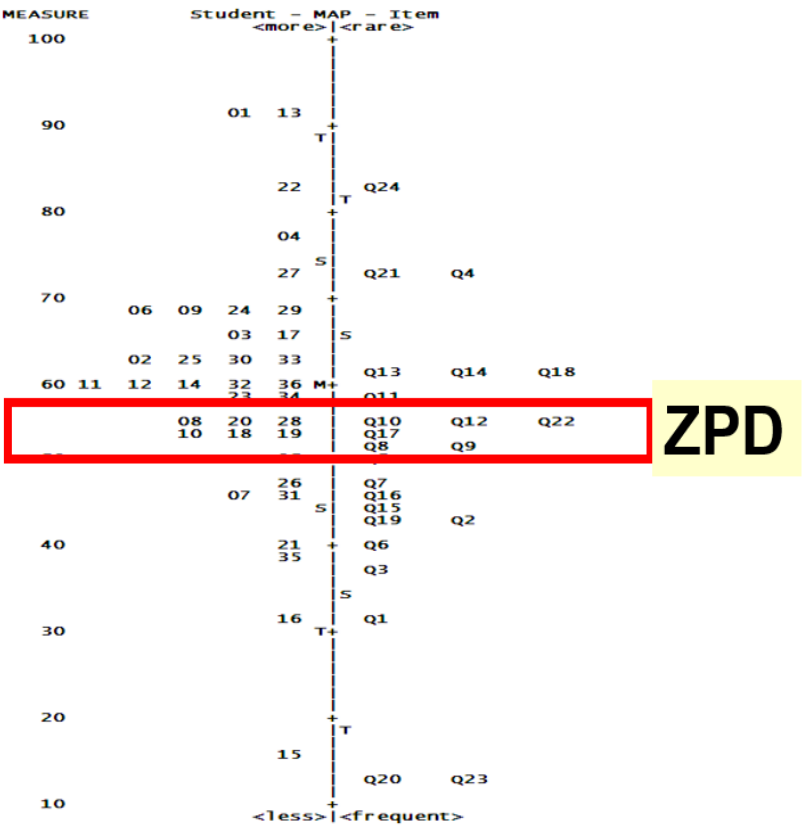


Figure 10. An example using a student-item map to identify the ZPD for an individual student

should focus on the contents and difficulty levels similar to these questions. Compared to assigning identical worksheets with a large number of items to all students, this approach is likely to be more efficient because the practice targets their ZPDs. Children will not waste their time on excessively easy items or get frustrated because of failure on too challenging items.

Of course, teachers can slightly extend or narrow the range of the ZPD according to different purposes. But the key point here, which is also at the core of assessment for learning, is identifying each student's ZPD, and guiding them to achieve their learning targets.

6.2. The PKMAP

While the variable map provides an overall picture for a group of students, the student diagnostic map (i.e., PKMAP) offers valuable information about the response pattern at the individual student level. Figure 11 is an example of the PKMAP, where *x.y* stands for the student's response *y* on item *x*. For example, "21.1" implies that the student had a correct answer (1) on item Q21, while "24.0" shows that the student gave an incorrect answer (0) to item Q24. The location of "xxx" indicates the overall person achievement measure of the student on the test. The red line above and below the student's location indicates the upper and lower bound of the ability estimate (ability estimate plus or minus one standard error: *bn* +/- *sn*). For items within the range between the two lines, it is reasonable to expect the student might answer them correctly (e.g., Q13 and Q14) or incorrectly (e.g., Q18) because those items' difficulties are close to the student's ability. The student's responses to items in the top right area and bottom left area are within Rasch model expectations, too. In other words, the student failed items with difficulty levels higher than the student's ability (e.g., Q4 and Q24) and succeeded on items with difficulty levels lower than the student's ability (e.g., Q1 and Q3). However, the student's responses to items in the top left area and bottom right area are more or less unexpected. The difficulty level of item Q21 is much higher than the student's ability, but the student got it correct. In contrast,

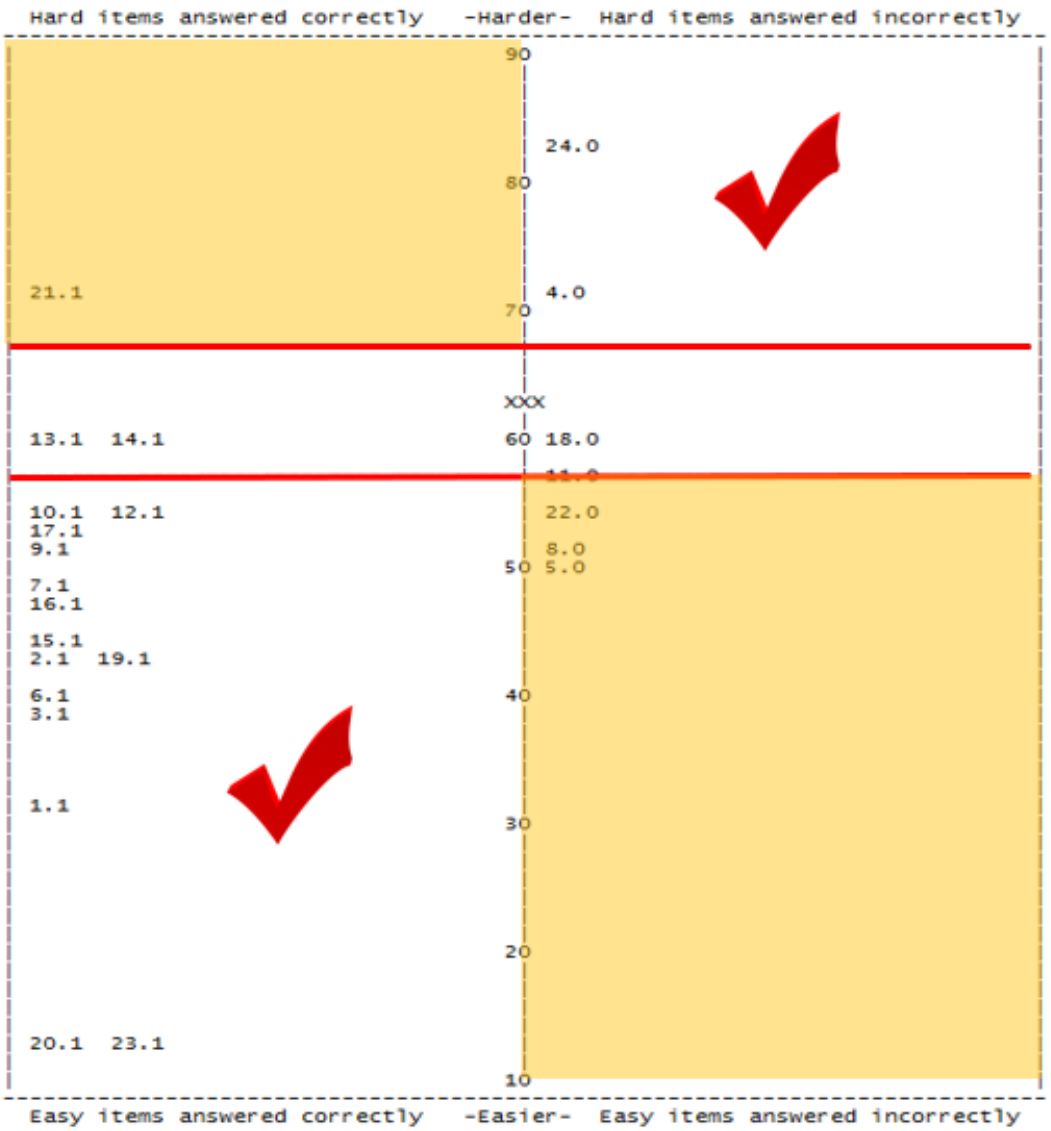


Figure 11. An example of PKMAP

the difficulty level of items Q22, Q5, and Q8 are lower than the student's ability, but the student got them wrong. These "warning messages" (esp. Q22, Q5, and Q8) alert the teacher that it would be worthwhile to investigate the reasons behind such unexpected responses.

In addition to facilitating teachers' instruction, the PKMAP has the potential to foster "assessment as learning" in which students agency in learning is emphasised. By interpreting the information from the PKMAP, students can have a nuanced understanding of their own assessment performance against the learning target. They can actively identify their learning needs and take appropriate actions to follow up their own learning. While this could happen in traditional ways of assessment analysis, the Rasch measurement offers more individualised and user-friendly information to facilitate its occurrence.

7. Conclusion

This article provides arguments for the applications of Rasch measurement in five areas (i.e., developing new instruments, creating short-form instruments, developing vertical scales, combining Rasch analysis and path analysis, and applications to classroom testing) with concrete examples. I do not intend to provide a comprehensive account of the applications

of Rasch measurement but to showcase the merits of the Rasch model in solving practical measurement problems in different contexts and highlight the principles researchers should attend to when applying the Rasch model. Researchers and practitioners are strongly encouraged to find other meaningful and promising areas to extend the application of Rasch measurement further and enact its benefits to research and practice.

Reference

Aryadoust, V., Tan, H.A.H., & Ng, L.Y. (2019). A Scientometric review of Rasch measurement: The rise and progress of a speciality. *Frontiers in Psychology*, *10*, 2197.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. (4th ed) London: Routledge.

Linacre, J. M. (2011). *Winsteps (Version 3.72.3)* [Computer Software]. Chicago: Winsteps.com.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749.

Yan, Z. (2018). The self-assessment practice scale (SaPS) for students: Development and psychometric studies. *The Asia-Pacific Education Researcher*, *27*(2), 123-135. https://doi.org/10.1007/s40299-018-0371-8

Yan, Z. (2020). Developing a short form of the self-assessment practices scale: Psychometric evidence. *Frontiers in Education*, 4, Article 153. https://doi.org/10.3389/feduc.2019.00153

Yan, Z., & Brown, G. T. (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. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., Brown, G. T., Lee, J. C. K., & Qiu, X. L. (2020). Student self-assessment: Why do they do it? *Educational Psychology*, *40*(4), 509-532. https://doi.org/10.1080/01443410.2019.1672038

Yan, Z., Lau, D. C. H., & Mok, M. M. C. (2013). A concurrent-separate approach to vertical scaling. In M. M. C. Mok (Ed.), *Self-directed Learning Oriented Assessments in the Asia-Pacific* (pp. 187-201). Springer.