Development and psychometric evaluation of the Generic Skills Teacher-Rating Scale for students with SEN in Hong Kong

F. T. F. Ye, F. Gao, L. Yang, C. L. Hsu & K. F. Sin

針對香港有特殊教育需要學生之共通能力測試量表（教師版）的開發及測評

葉天放、高鳳展、楊蘭、許嘉凌、冼權鋒

Abstract

Research reveals that generic skills are essential for supporting students to not only finish school education but also transition to the workplace and adult life. However, a comprehensive literature review found limited studies have been done to develop an assessment tool for examining the generic skills of students with special educational needs (SEN). Based on two sub-studies, the current study aimed to develop a short form of the Generic Skills Teacher-Rating Scale (GSTS) to assess six generic skills (i.e., Collaboration， Communication, Problem-solving, Self-management, Information Technology, Critical thinking) of SEN students. A total of 231 Hong Kong students from six special schools (including primary and secondary levels) participated in the two studies. Both CFA and Rasch analyses support a six-factor solution of the GSTS with reasonably good model fit indices. Researchers and educators may view the GSTS as an assessment tool, adding more information on SEN students' generic skills beyond classroom settings to understand the daily-life functioning of students in special education and inclusive education settings. Implications of the results and future directions are also discussed.

研究表明，共通能力對於支持學生完成學業以及過渡到未來工作和日常生活至關重要。 然而，一項全面的文獻回顧發現，有關評測有特殊教育需要 (SEN) 學生的共通能力的評估工具方面進行的研究非常有限。本研究（基於兩項子研究）旨在開發一項針對 SEN 學生的共通能力教師評定量表 (GSTS) 的簡短表，以評估六種共通能力，即協作能力、溝通能力、解決問題能力、自我管理能力、運用信息科技能力、批判性思考能力。 來自六所特殊學校（包括小學和中學）的共二百三十一名香港學生參與了這兩項子研究。CFA 和 Rasch 分析都支持 GSTS 的六因素解決方案，表明該量表具有相當好的模型擬合指數。 研究人員和教育工作者可以將 GSTS 視為一項評估工具，用於在課堂環境之外了解有關 SEN 學生共通能力表現，以助提高特殊教育和全納教育環境中學生的日常生活功能。本文還討論了結果的含義和未來的方向。

Keywords:

Generic Skills	共通能力
Special Education	特殊教育
Scale Development	量表開發
Psychometric Evaluation	量表測評

Introduction

Generic Skills

Generic skills, also known as "employability skills", refer to the skill sets that can be applied across occupations and daily-life scenarios (Chan & Fong, 2018). These skills are crucial to the whole-person development of students and have been attracting growing attention in higher education around the world (Cheng et al., 2018), as such skill-based pedagogy provides valuable training to help students transit into higher education and different disciplines (Star & Hammer, 2008). For instance, students equipped with generic skills would have better employment prospects, could transfer skills across different jobs, and process the ability to be lifelong learners (Freudenberg et al., 2011). In response to the global trend of nurturing graduates with holistic competency and career development, institutions and schools worldwide have been developing and implementing learning activities targeting students' generic skills. In Hong Kong, a great emphasis on integrating generic skills into curriculum development has occurred for over two decades (Education Bureau, 2015). For effective integration and implementation of generic skills in the curriculum, the Curriculum Development Council of Hong Kong Education Bureau (2015) identified nine essential skills and highlighted in the curriculum framework. These skills include communication skills, mathematical skills, information technology skills, critical thinking skills, creativity, problem-solving skills, self-management skills, self-learning skills, and collaboration skills. The nine generic skills provided schools with a rich and flexible framework to infuse these skills into curriculum modes.

Generic skills are part of students' life skills. As Yuen et al. (2010) highlighted for schools, they "strive to equip their students with a set of generic transferable skills that enable them to take on various functions and life roles, such as learners, friends, workers, parents and citizens" (p. 296). To equip students with special educational needs (SEN) with generic skills is more crucial to their life skills development and career pathways as compared to their counterparts without SEN (Yang et al., 2020). Similarly, Solberg et al. (2020) also argued that assisting students with SEN to engage in career learning requires recognizing their strengths and empowering them with life skills. The goal of quantifying and emphasizing generic skills is to equip students with skills or attributes beyond disciplinary content knowledge, which can be broadly applied across different contexts in daily life. Some researchers have identified skills commonly seen as critical to students' development. For example, critical thinking, problem-solving, interpersonal skills, communication, and information management skills, and so on (Barnett et al., 2005). However, assessment tools of generic skills with good psychometrical properties are needed to integrate generic skills development to curriculum adaptation and accommodation in special education and inclusive education.

The Assessment of Generic Skills

The typical method of assessing generic skills involves self-assessment and well-informed professional judgment by raters as well as teams of experts in the field (Gibb, 2004). Summarized in Curtis and Denton (2003), four approaches could be applied for the purpose of assessment: targeted instruments, holistic judgment, student portfolios, and work experience. Curtis and Denton (2003) also discussed the key features, for example, the validity and reliability of the assessment tools, and suggested that generic skills should be explicitly incorporated in training programs associated with the assessments. However, conducting self-assessment among students with SEN would encounter validity-related problems and be practically difficult to operate in classrooms. Thus, an easy-to-use measurement tool is required for teachers in special education.

Despite awareness of the importance, generic skills have not received enough attention in special education. To date, there is not yet a measurement tool developed for the purpose of quantifying generic skill learning outcomes among students with SEN. In response to the need for generic skills measurement tools in special education settings, the current study aimed to develop the Generic Skills Teacher-Rating Scale as a convenient and reliable measurement of generic skill learning outcome indicators for students with SEN in Hong Kong.

The current project prioritized six out of nine generic skills in measurement development, even though the nine generic skills are the expected universal learning outcome for all students at the curriculum planning level (Curriculum Development Council, 2001 & 2015). As the current project aimed to develop activity-based learning techniques and virtual reality technology-based training programs for students with SEN, the six skills were chosen based on the availability of the tasks, as well as the universality across mild, medium, and severe level special educational needs. Furthermore, in line with the program curriculum, the definitions of the six generic skills were reviewed by special education professionals in Hong Kong and operationalized into measurable items. Mathematical skills, self-learning skills, and creativity will be implemented in future studies and training programs that utilize different learning and teaching strategies in classroom scenarios. As a result, the current measurement tool included six out of nine generic skills defined as follows:

- Communication skills: the abilities to express himself / herself through all possible means and understand basic instructions.
- Collaboration skills: the ability to work with others in a team.
- Critical thinking: drawing out meaning from available information and making own judgments.
- Problem-solving skills: the ability to resolve daily life difficulties, sometimes with the assistance of others.

- Self-management Skills: the ability of independence, including the ability to maintain emotional stability and exercise self-discipline.
- Information Technology (IT) skills: the utilization of IT equipment to do simple searching and sharing in IT interfaces.

The Rasch Modeling

To have a better understanding of the qualities among a large number of test items, Rasch analysis was employed to assess the psychometric properties of the Generic Skills Teacher-Rating Scale (GSTS). Rasch analysis assumes that the items in the measurement mandate a unidimensional latent construct/trait, and the items and subjects can be fitted along the trait continuum according to their varying level of difficulty level/ability level (Bond & Fox, 2007). Furthermore, it can provide item-level evaluations through item fit statistics. Therefore, using Rasch analysis, firstly, allows us to evaluate the extent to which the unidimensional scale was constructed at the item and person levels. That is, it provides more fit indices to evaluate the measurement tool's quality. If the items fit adequately, the difficult items will be endorsed by fewer people, while more people will endorse the easy items. Similarly, for person fit, individuals with low-level latent traits tend to endorse easy items rather than difficult items. Secondly, the Rasch model is an objective model compared to other item response theory models and classical testing theory models (Bond & Fox, 2007). Its model fit indices evaluate the discrepancy between empirical data and the hypothesized model without adjusting the model to compensate the model misfit. Particularly, it is useful for selecting well-fitted items from the item pool (Andrich & Marais, 2019; Bond & Fox, 2007). As a result, the Rasch model has gained growing attention and has been a popular choice of psychometric evaluation tools in the psychological literature (Andrich & Marais, 2019; Boone et al., 2013).

The Current Studies

The current studies were conducted in parts of evaluation studies in a grand project that aims at facilitating experiential learning and promoting generic skills to SEN students. Targeting six specific generic skills, the project utilized activity-based learning techniques and virtual reality technology, and developed and conducted a full range of training packages with assistance from stakeholders. On this basis, the studies described here intended to develop and evaluate a new teacher-rating instrument for educators in special education to measure SEN students' generic skills learning outcomes. To achieve this aim, two consecutive studies were conducted among students from six special schools in Hong Kong. In Study 1, a large item pool was generated by a group of experts, and the items were rated by teachers from special schools. Then, Rasch analysis was used for evaluating and selecting qualified items. Next, in Study 2, the selected items from Study 1 were rated on a different sample by the teachers and were evaluated based on item functioning. The Rasch model was used to evaluate the items' qualities in both Study 1 and Study 2. In

addition, confirmatory factor analysis was used as supplemental evidence to examine the test-level examination of the finalized scale.

Study 1 Psychometric Evaluation of The Initial Item Pool of 332 Items

The purpose of Study 1 is to examine the psychometric property of the initial item pool of the GSTS and select adequate items for item calibration in the subsequent study.

Method

Item Development

In Study 1, the research team first formed a group of experts that consisted of experienced teachers, educational psychologists, specialists in special education, psychometricians, and other researchers. The group conducted a number of school visits, class observations, and interviews with teachers to construct the measurement framework.

Through discussions with these experts, the initial item pool was developed through an iterative process based on the program curriculum and professional input from teachers. The goal was to generate an over-inclusive item pool that covers as many aspects of the generic skill domains as possible. To ensure items had good face validity, all items were distributed to a research group of 17 experts (10 teachers, 6 school principals, and 1 education psychologist) for further evaluation. These experts provided detailed suggestions regarding item content, wordings, and general methodological feedback on the assessment tool. Based on these suggestions, items were further modified through discussions with researchers. Eventually, the GSTS item pool contained 332 items in 6 sub-scales measuring corresponding skill domains: collaboration (48 items), communication (48 items), problem-solving (64 items), self-management (72 items), IT skills (44 items), and critical thinking (56 items). All items were assessed by asking teachers to indicate to what extent a student exhibits specific behaviors related to generic skills on a 5-point Likert scale (from 1 = never to 5 = always).

Participants

One hundred and seventy-six students from six special schools (including primary and secondary level) in Hong Kong participated in the study. Students included 121 males and 55 females, and aged from 6 to 21 (Mean = 11.5, Standard Deviation = 3.8). Among them, 61 were with mild level intellectual disability, 60 with moderate level intellectual disability, and 55 with severe level intellectual disability. Detailed demographic information is shown in Table 1. Informed consent from students and their parents was obtained prior to data collection. If they did not want to be graded, parents and students were allowed to opt-out at any time during the study without consequence. Before submitting for data analysis, all the data were

anonymously coded with a unique code generated for each student to protect their confidentiality.

Table 1 Demographic information of SEN students in Study 1

SEN Type	SEN level			Gender		Age		Total
	Mild	Moderate	Severe	Male	Female	Mean	SD	
ID	40	44	37	85	36	11.5	3.98	121
ID+ASD	18	14	-	23	9	11.8	3.9	32
ID+PD	-	0	12	6	6	11.1	3.1	12
ID+VI+PD	-	0	5	3	2	9.4	0.9	5
ID+Down Syndrome	-	2	1	3	-	12.7	3.2	3
ID+ADHD	2	-	-	-	2	15.0	0	2
ID+others	1	-	-	1	-	12.0		1
Total	61	60	55	121	55	11.5	3.8	176

Note. ID=Intellectual Disability; ASD=Autism spectrum Disability; PD=Physical disability; VI=Visual Impairment; HI=Hearing Impairment; and ADHD=Attention Deficit/ Hyperactivity Disorder.

Data Collection

Teachers who were familiar with their students' conditions were invited to provide ratings for each student. As a result, students were rated by their corresponding head teachers in each school, and the teachers were all verified special educators and had years of teaching experience. Data collection was conducted within two months. Teachers were first briefed by the researchers regarding the rating standards and the meanings of items, and then the assessment tool was distributed to them in the online survey format. Teachers were required to complete the assessment on computers.

Analytical Strategies and Item Selection

The team utilized the Rasch rating scale model (RSM; Andrich, 1978) to examine the psychometric properties of each GSTS sub-scale and for item selection. Analyses were conducted using Winsteps 4.3.1 (Linacre, 2018a). Items were discarded or retained based on the model fit described below.

Dimensionality. Utilizing Rasch analysis requires the items used in the model to reflect a unidimensional construct. Thus, principal components analysis of residuals (PCAR) was conducted for each sub-scale before examining fit indices. As suggested by Linacre (2018b), a measurement can be regarded as unidimensional if the variance explained by the measures is substantial (e.g., more than 40%) and the variance explained by the first contrast is negligible (e.g., less than 5%).

Rating Scale Functioning. The reliability of the instrument was evaluated by item separation and person separation, which indicates the spread of items or persons relative to the standard errors. Greater separation means greater reliability. A person separation value greater than 2 suggests the instrument can sufficiently distinguish high and low performers, and an item separation value greater than 3 suggests the sample is sufficient to confirm the item hierarchy (Linacre, 2018b).

For item selection, Rasch infit and outfit were examined in terms of the mean square residual (MNSQ) with the criterion of $.6 \leq \text{MNSQ} \leq 1.4$ that recommended for the items' qualities as an acceptable fit range (Wright et al., 1994). Infit gives more weight to individuals who are close to item difficulty; therefore, it is less sensitive to outliers than outfit (Bond & Fox, 2007). Thus, when examining misfit items, emphasis was put on infit statistics. For developing an instrument, we also expected items to be better behaved than persons (Wright et al., 1994), so the emphasis was put on item fit instead of person fit. In addition, we examined the item-total correlation for each item. Items with an item-total correlation value that is greater than .3 will be considered as better reflecting the target construct (Linacre, 2018b).

Results and Discussion

The dimensionality of the item pool was examined separately for each sub-scale. The results of PCAR showed that, for collaboration, 75.7% of the variance was explained by measures with only 3.8% unexplained in the first contrast. Thus, the collaboration scale was considered unidimensional. Similarly, for communication, 80.8% was explained by measures, and 4.2% in the first contrast; for problem-solving, 80.5% was explained by measures, and 3.7% in the first contrast; for self-management, 76.6% was explained by measures, and 3.0% in the first contrast; for IT, 81.2% was explained by measures, and 4.0% in the first contrast; and for critical thinking, 81.1% was explained by measures, and 3.0% in the first contrast. In summary, all sub-scales satisfied the assumption for unidimensionality. Besides, initial screening showed that all items had item-total correlation values above .45, which was beyond the required threshold of .3. Thus, all items were retained for item fit analysis.

Table 2 presents the summary of reliability indices and item fit statistics of the item pool. Based on the results, 132 items were discarded due to inadequate fit or item measure redundancy (overlapping item measure). Regarding face validity and content importance, identified misfit items were further inspected and discussed by researchers before final deletion. As a result, 200 items were retained in the initial version.

Overall, starting from a pool of 332 items, Study 1 yielded a set of 200 items based on theoretical and practical concerns. As an initial version of the measurement

of the generic skills, these selected items demonstrated adequate item fit and were submitted for further analysis in Study 2.

Table 2 Model and item statistics in Study 1

Sub-scale	Initial Item Pool	Separation		Item Infit MNSQ	Item Outfit MNSQ	Selected Items
		Item	Person			
Collaboration	48	9.31	7.36	0.54 ~ 1.52	0.51 ~ 2.45	28
Communication	48	10.18	6.89	0.47 ~ 1.83	0.39 ~ 3.18	36
Problem-solving	64	8.69	9.95	0.56 ~ 2.54	0.49 ~ 3.38	37
Self-management	72	7.31	8.37	0.48 ~ 1.88	0.42 ~ 1.78	42
IT	44	9.27	6.08	0.44 ~ 2.90	0.26 ~ 3.41	12
Critical thinking	56	9.02	8.23	0.44 ~ 1.91	0.51 ~ 2.62	45

Note. MNSQ = Mean Square Residual.

Study 2 Psychometric Evaluation of The Selected 200 Items

In Study 2, the selected items in Study 1 were submitted to a retest with a larger sample in a similar setting to refine and shorten the scale and evaluate its factorial structure.

Method

Participants

In Study 2, the team administered the 200 items to a sample of SEN students after four months of the first data collection. In total, 231 students were rated by teachers, including 93 students who participated in Study 1, and 138 students who were newly recruited. The demographic information is shown in Table 3. The final sample contained no missing data.

Analytical Strategies and Item Selection

Rating Scale Functioning. Like Study 1, the team used the RSM (Andrich, 1978) to evaluate the large set of items' psychometric properties. Then, following the recommendations from Wright (1994), the team prioritized investigating and removing underfitting items with high randomness (high MNSQ) rather than overfitting items that are too predictable (low MNSQ), because high MNSQ items would distort or degrade the measurement. To effectively shorten the scale, items were evaluated with high stake standards, $0.8 \leq \text{MNSQ} \leq 1.2$. Similar to Study 1, Rasch analysis was performed for each sub-scale. Each time, the most misfit item was identified, checked, and deleted, and the remaining items were submitted for a retest. This process was repeated until the sub-scale functioning satisfied the requirement.

Table 3 Demographic information of SEN students in Study 2

SEN Type	SEN level			Gender		Age		Total
	Mild	Moderate	Severe	Male	Female	Mean	SD	
ID	38	33	65	82	55	12.09	3.40	137
ID+ASD	32	20	1	41	12	12.11	3.30	53
ID+PD	-	-	10	3	8	13.09	4.48	11
ID+VI+PD	-	-	5	4	1	9.40	.89	5
ID+Down Syndrome	1	7	1	4	5	10.78	4.29	9
ID+ADHD	5	-	-	2	3	15.40	2.07	5
ID+HI	1	-	2	1	2	12.67	5.03	3
ID+others	4	1	3	7	1	10.75	2.38	8
Total	81	61	87	144	87	12.06	3.44	231

Note. ID=Intellectual Disability; ASD=Autism spectrum Disability; PD=Physical disability; VI=Visual Impairment; HI=Hearing Impairment; and ADHD=Attention Deficit/ Hyperactivity Disorder.

Differential Item Functioning. To ensure the measurement to be invariant across gender, we examined differential item functioning (DIF) for each selected item. If an item presents significant DIF, it indicates its logit position is biased towards one gender over the other. DIF occurs when an item has different item measures across groups, and its corresponding significance test (Welch t-test) shows a *p*-value less than .05. However, as suggested by Linacre (2018b), a difference (DIF contrast) that is less than .43 can be considered as small and negligible.

Additional Criterion. Items with high similar difficulty and content were identified, evaluated, and removed. Additionally, the category probability curve (CCC) and item characteristic curve (ICC) of each item was visually inspected. If a CCC shows clear and ordered thresholds of the item category probability, and an ICC demonstrates a match between empirical probability and hypothesized model, it indicates the item has well-functioned response options (Bond & Fox, 2007). We also examined the Wright map, in which persons and items were plotted on the same continuum according to the ability and difficulty estimates to understand the person-item relations.

Reliability. For the short form, the internal reliability of each sub-scale was examined. Apart from the person and item separation provided by Rasch analysis, we reported the Rasch reliability as the supplementary index. Using the identical sample in both studies, we also calculated test-retest reliability for each sub-scale to examine the instrument's stability across two-time points. The test-retest reliability was calculated using psych (Revelle, 2021).

Confirmatory Factor Analysis. The team conducted confirmatory factor analysis (CFA) to evaluate the six-factor structure of the GSTS short form. The CFA model was estimated with robust maximum likelihood (MLR) estimator using lavaan in R (R Core Team, 2020; RStudio Team, 2018). Model fit was evaluated based on conventional cut-offs: the model will be considered as an adequate approximation to the data when CFI, TLI > .9, RMSEA, SRMR < .08; it will be considered as good when CFI, TLI > .95, RMSEA, SRMR < .06 (Hu & Bentler, 1999; Kline, 2016).

Results

Prior to Rasch analysis, the team examined data for the assumption of unidimensionality. Again, all sub-scales satisfied the requirement of unidimensionality, with more than 70% of the variance explained by the measure and less than 5% of the variance explained by the first contrast. All items had item-total correlation values greater than .8.

In general, the communication, problem-solving, critical thinking and IT sub-scales were rated on average lower than collaboration and self-management sub-scales. The mean person measure of sub-scales were: collaboration = -.56, communication = -1.18, problem solving = -1.69, self-management = -.33, IT = -3.48, critical thinking = -1.54.

Following the iterative process described above, in total 152 misfit items were discarded across six sub-scales. Additionally, 6 items exhibit substantial (DIF contrast > .48) and significant (p < .05) gender DIFs. The remaining 42 items comprised the final GSTS short form. The item CCCs showed that all items had ordered thresholds for category probability. The CCC of a sample item in the critical thinking sub-scale was demonstrated in Figure 1. The ICCs of each sub-scale also supported the proper functioning of the scale, showing that the empirical curves match the theoretical curves reasonably well. The Wright map suggested that the communication, problem-solving, critical thinking and IT sub-scales were relatively difficult, as most items were intended for high-ability SEN students; meanwhile, the collaboration and self-management sub-scales had normally distributed item difficulty as well as person estimates, indicating these items were suitable for the current sample. Figure 2 presents all ICCs and the Wright map for each sub-scale.

The GSTS short form demonstrated excellent reliability (Table 4). Of all six sub-scales, the person separation ranged from 2.78 to 4.23, and the item separation ranged from 2.90 to 10.65. The reliabilities of the six sub-scales ranged from .89 to .95, indicating the instrument had good stability. Additionally, the ratings of identical subjects between the two studies showed that the test-retest reliability of the six sub-scales ranged from .71 to .92, indicating the instrument had good stability over time.

Table 4 Model and Item Statistics in Study 2

Sub-scale	Short Version	Separation		Item Infit MNSQ	Item Outfit MNSQ	Reliability
		Item	Person			
Collaboration	10	10.65	4.23	0.89 ~ 1.10	0.80 ~ 1.14	0.95
Communication	6	8.74	4.01	0.89 ~ 1.17	0.67 ~ 1.08	0.94
Problem-solving	6	2.90	3.92	0.85 ~ 1.10	0.81 ~ 1.12	0.94
Self-management	6	8.66	3.59	0.83 ~ 1.09	0.81 ~ 1.09	0.93
IT	5	3.55	2.78	0.90 ~ 1.07	0.85 ~ 1.19	0.89
Critical thinking	9	5.13	4.21	0.74 ~ 1.15	0.75 ~ 1.29	0.95

Figure 1 The Category Probability Curve of a sample item (Item 255) in critical thinking sub-scale

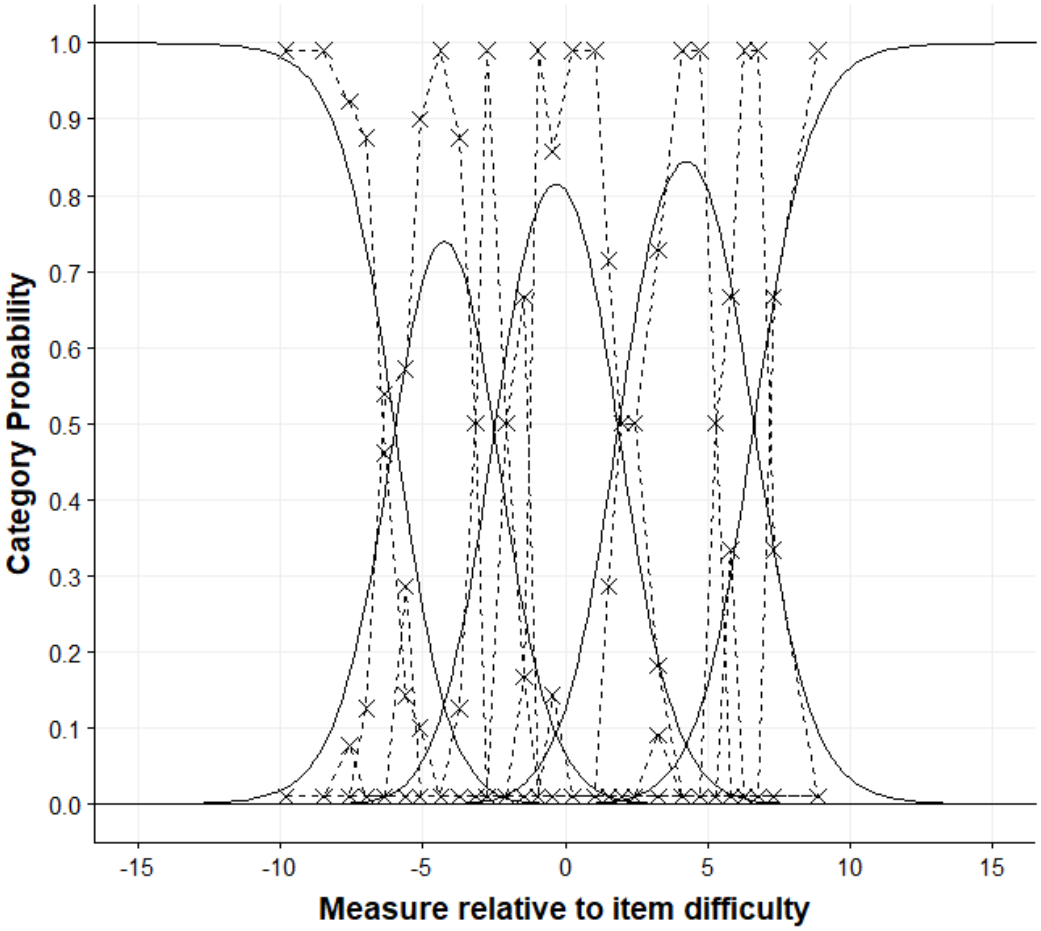
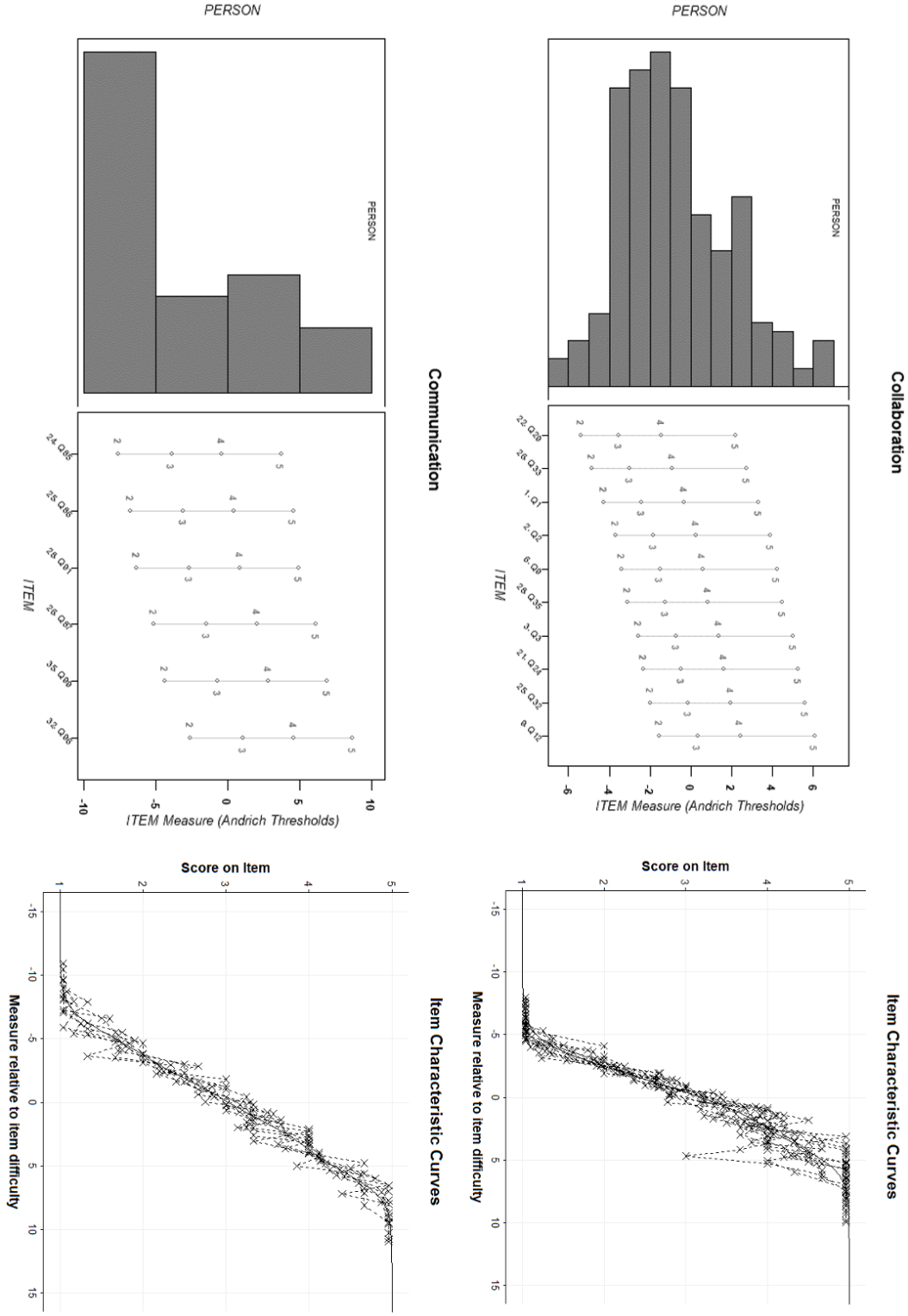
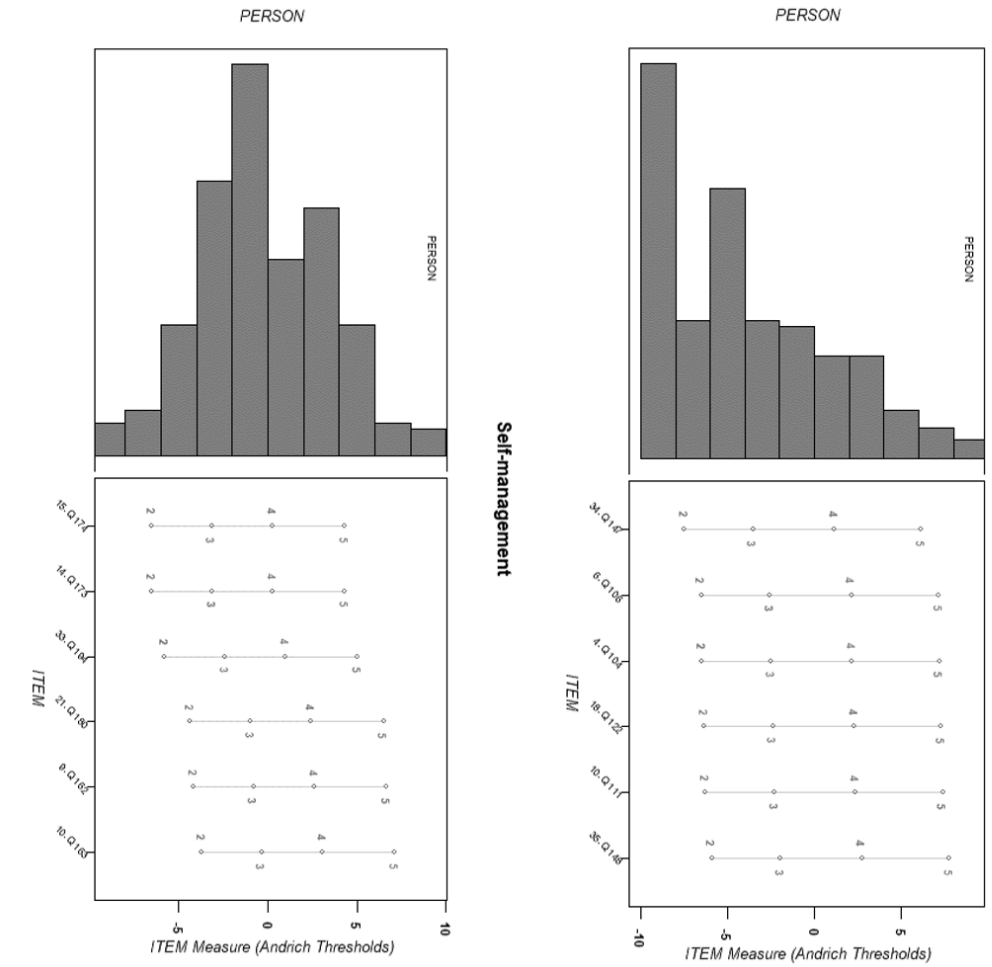
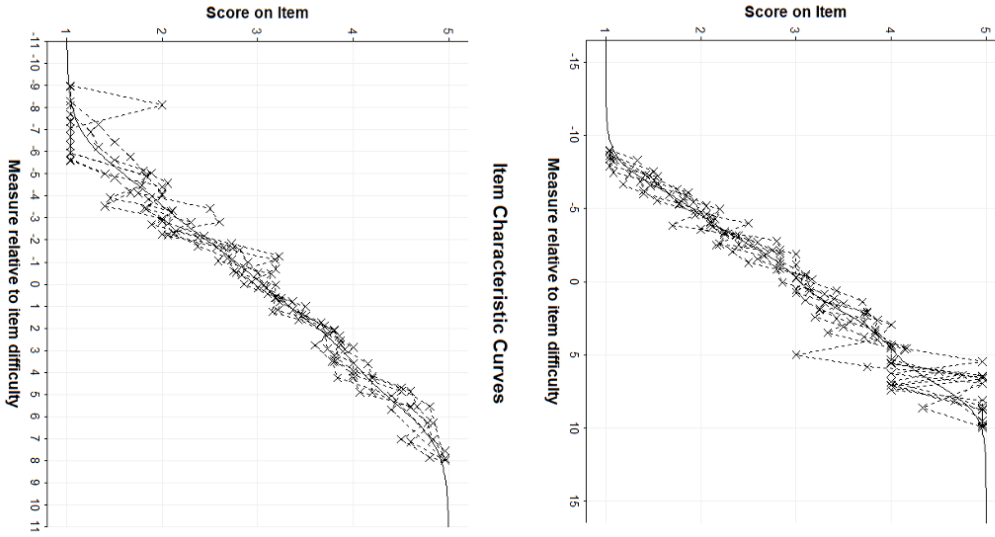
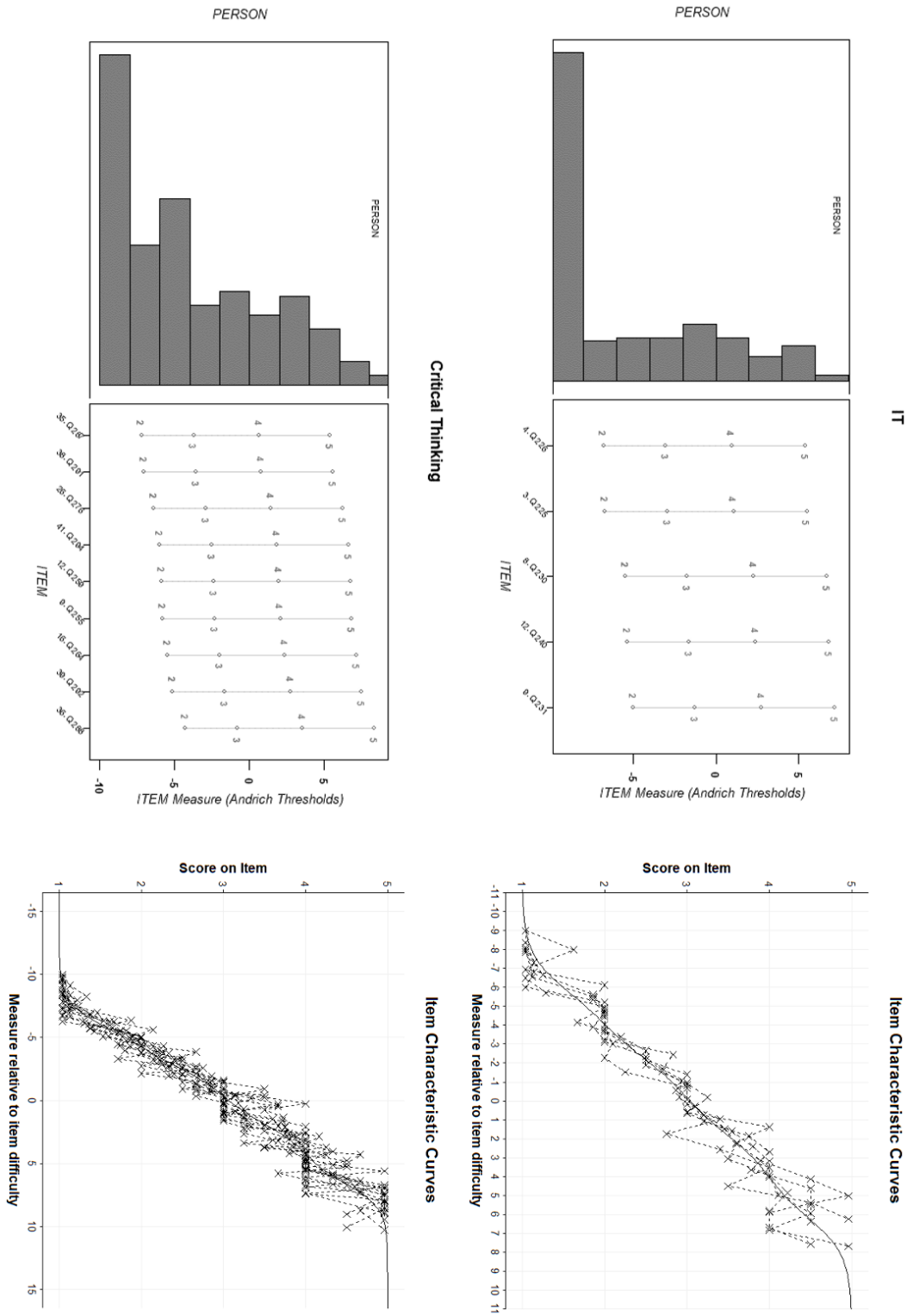


Figure 2 Item Characteristic Curves and Wright Maps for the Six Sub-Scales




Problem Solving
PERSON
ITEM
ITEM Measure (Andrich Thresholds)
Self-management
PERSON
ITEM
ITEM Measure (Andrich Thresholds)


Item Characteristic Curves
Score on Item
Measure relative to item difficulty
Item Characteristic Curves
Score on Item
Measure relative to item difficulty


IT
PERSON
PERSON
ITEM
4. Q226
3. Q225
8. Q230
12. Q240
9. Q231
ITEM Measure (Andrich Thresholds)
Critical Thinking
PERSON
PERSON
ITEM
35. Q206
36. Q207
26. Q276
41. Q204
12. Q236
9. Q235
16. Q264
39. Q202
38. Q208
ITEM Measure (Andrich Thresholds)
Item Characteristic Curves
Score on Item
Measure relative to item difficulty
Item Characteristic Curves
Score on Item
Measure relative to item difficulty

The final version of the scale was submitted to CFA. In the measurement model, items in six sub-scales were loaded onto their corresponding latent factors, the latent factor variance was fixed to 1, and the covariance values among six latent factors were free to estimate. The results show that the model fit the data well, Robust $\chi^2(804) = 2067.70$, CFI = .910, TLI = .904, RMSEA = .088, SRMR = .048. All factor loadings were significant, $p < .001$, and ranged from .72 to .96. Given that RMSEA only exceeded the cutoff value for a small amount, and other model fit indices were all above the threshold, no modification was employed, and the CFA model was retained. Thus, the team concluded that the six-factor model was an adequate approximation to the data. However, the inter-factor correlations among the six generic skills ranged from .72 to .98, indicating some of the factors were considered highly similar and lack of discriminant validity (Kline, 2016).

Additionally, the team conducted repeated measures analysis of variance (ANOVA) to describe the general profile of the six generic skills among SEN students. In the ANOVA, the mean scores of the six sub-scales were entered as within-subject factor, and the results showed a significant main effect, $F(5, 1150) = 211.74$, $p < .001$, $\eta^2 = .48$, indicating that there were significant differences in six generic skills. Post Hoc tests with Holm correction showed that self-management ($M = 2.95$) was higher than collaboration ($M = 2.73$), $d = .35$, $p < .001$, communication ($M = 2.19$), $d = 1.21$, $p < .001$, problem-solving ($M = 2.16$) d = 1.26, $p < .001$, IT ($M = 1.81$), $d = 1.80$, $p < .001$, and critical thinking ($M = 2.09$), $d = 1.35$, $p < .001$. Collaboration was higher than communication, $d = .86$, $p < .001$, problem-solving, $d = .90$, $p < .001$, IT, $d = 1.45$, $p < .001$, and critical thinking, $d = 1.0$, $p < .001$. Communication was higher than IT, $d = .59$, $p < .001$, but not significantly different from problem-solving $d = .05$, $p = .47$, or critical thinking, $d = 14$, $p = .09$. Problem-solving was higher than IT, d = .54, p < .001, but not significantly different from critical thinking, $d = .10$, $p = .29$. Lastly, critical thinking was also higher than IT $d = .45$, $p < .001$.

Discussion

The current paper has presented an overview of two studies in which the team developed, shortened, and evaluated the psychometric properties of the GSTS using Rasch modeling and CFA. GSTS is an easy-to-use instrument designed for teachers to assess the six generic skills among SEN students in Hong Kong. In Study 1, an over-inclusive item pool was generated by conducting observations and reviewing literature in the field. Then, these items were rated among 176 SEN students in six special schools and were tested rigorously using Rasch analysis. Next, in Study 2, we selected 200 well-fitted items and conducted a second wave of data collection among 231 SEN students. Based on Rasch analysis, DIF tests, and CFA, the final version of GSTS retained 42 items as the short form.

Although the length of GSTS was greatly reduced, the scale demonstrated good construct validity, reliability, and stability compared to the original item pool. As were shown in the CCCs and ICCs, the empirical probability curves were closely matched with the modeled curves. Thus, the items in the final version functioned well in each sub-scale. These items were also DIF-free items with good test-retest reliability, which presented its potential usefulness in assessing generic skills among SEN students in classroom settings. While these items were responded to as intended by the raters, the proportion of the students with ratings in the upper categories was low, especially in the communication, problem-solving, IT, and critical thinking sub-scales. These poorly targeted items indicated that students in the current study had low performance in the four generic skill domains. On the one hand, this phenomenon suggests items in these four sub-scales are too difficult and should be calibrated in future studies; on the other hand, the results highlight the importance of providing training resources and promoting these skills among SEN students.

The final version of the scale yields a six-factor solution as intended with reasonably good model fit indices. However, the standardized inter-factor covariances between communication and problem-solving, communication and critical thinking, IT and critical thinking, problem-solving, and IT were above .9, indicating students' performance on these skills were perceived to be closely related and suggesting that these sub-scales demonstrated insufficient discriminant validity. Feedback from teachers revealed that most of the SEN students were considered to have a consistent performance, which means low ratings in one skill usually indicates low ratings in other skills, though the items were assessing different behaviors. Thus, a clear explanation cannot emerge offhand, and further study is needed.

The aim of developing a short form of the GSTS is to enhance its usability in classroom settings. In addition, researchers and educators may view the GSTS as a complementary tool, adding additional information on SEN students' generic skills beyond classroom settings to understand the daily-life functioning of students in special education and inclusive education.

Implications

Implications to educational assessment: This study has several implications. First, based on the Rasch analysis and CFA, this study systematically tested the item validates and factorial structures of GSTS, which provide the basic framework for future studies in terms of convergent validity and reliability. Second, the results of this study revealed that the final version of the GSTS (short form, 42 items) has high internal reliability and a clear six-factorial structure. The GSTS short-form, compared to the original scale (140 items), could be used as a parsimonious tool for frontline teachers to quickly check SEN students' development of the six generic skills in classrooms of special education and inclusive education.

Implications to instructional practices: Timely assessment of SEN students' development of six generic skills would contribute to effective instructional adjustments or improvements to enhance one or more generic skills students are struggling to develop.

The relatively low scores of the six generic skills measured in the current project suggested that promoting their generic skills is of importance to facilitate the success of the social integration of students with SEN. In school settings, research has suggested that collaboration is an effective approach to promoting academic integration and improving adjustment, peer acceptance, and group unity among students with SEN in inclusive classrooms (Turnbull et al. 2004). In the post-tests, the improved communication and collaboration skills indicate that the generic skill training program provides a learning environment in which both specific and general graduate qualities can be fostered. Besides, communication and collaboration skills also reflect the identified needs for job requirements in modern society. Equipping students with adequate and sufficient interpersonal skills would help them communicate with others and listen to them in more sophisticated ways, and therefore provide them with more opportunities in the job market and higher life quality. On the other hand, the overall weak performance of IT, problem-solving, and critical thinking skills in the current project highlighted the importance of providing relevant resources in special education. These skills are interrelated and essential for students with SEN to develop self-help skills (Fitzgerald & Koury, 1996; Norman et al., 2001). Previous research also suggested that technology-assisted instruction can improve these skills (Fitzgerald & Koury, 1996; Cheng & Lai, 2020). Therefore, the current project offered valuable insights into applying modern technologies in special education classrooms.

Improving useful instructional designs and practices to support SEN students' development of generic skills may also involve other stakeholders aside from teachers. The whole school approach adopted in Hong Kong (Hui, 2002) would be beneficial to support "teachers in their guidance role of collaborating with guidance professionals to conduct guidance curriculum, student individual planning" to help students with diverse abilities (Yuen et al., 2010, p. 307). Given the various challenges and difficulties that teachers may meet in developing SEN students' six generic skills, the recent development of technology-enhanced approaches to teaching (e.g., virtual reality [VR] and augmented reality [AR]) can also be considered in designing effective instructional practices to support SEN students' learning and enhance their generic skills (e.g., Badilla-Quintana et al., 2020; Cascales-Martínez et al., 2016).

Limitations and Future Directions

The GSTS was developed and intended to measure generic skills among SEN students in classroom settings. Although the results supported its psychometric properties, the scale needs additional research to evaluate its convergent and discriminant validity. It is noted that some of the sub-scales exhibited relatively high correlations in the current studies. Such a phenomenon indicates that future research should look into greater details of the behavior indicators and consider modifying item content or combining factors.

Given that the ratings were provided by limited raters, the inter-rater reliability of the scale should be further tested with criterion-related variables. Although prior to data collection, training and discussions were conducted among raters, and the raters were trained by professionals with years of special education experience, different raters may still exhibit different levels of bias and tolerance in judgment. More research is needed to establish criterion-related validity and compensate for the lack of variance in the current studies. In addition, more research should be done to test whether the overall low scores on communication, problem-solving, critical thinking and IT sub-scales are due to demanding items or raters' misperceptions of the items. Lastly, the GSTS should be tested in larger diverse samples in different special education settings to explore its generalizability.

References

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer.

Badilla-Quintana, M. G., Sepulveda-Valenzuela, E., & Salazar Arias, M. (2020). Augmented reality as a sustainable technology to improve academic achievement in students with and without special educational needs. *Sustainability*, *12*(19), 8116.

Barnett, R., & Coate, K. (2005). *Engaging the curriculum in higher education*. Buckingham: The Society for Research into Higher Education.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences, 2nd ed.* Lawrence Erlbaum Associates Publishers.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *rasch analysis in the human sciences*. Springer Science & Business Media.

Cascales-Martínez, A., Martínez-Segura, M. J., Pérez-López, D., & Contero, M. (2016). Using an augmented reality enhanced tabletop system to promote learning of mathematics: A case study with students with special educational needs. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(2), 355-380.

Chan, C. K. Y., & Fong, E. T. Y. (2018). Disciplinary differences and implications for the development of generic skills: A study of engineering and business students' perceptions of generic skills. *European Journal of Engineering Education*, *43*(6), 927–949. https://doi.org/10.1080/03043797.2018.1462766

Cheng, M. W. T., Lee, K. K. W., & Chan, C. K. Y. (2018). generic skills development in discipline-specific courses in higher education: A systematic literature review. *Curriculum and Teaching*, *33*(2), 47–65. https://doi.org/10.7459/ct/33.2.04

Cheng, S.-C., & Lai, C.-L. (2020). Facilitating learning for students with special needs: A review of technology-supported special education studies. *Journal of Computers in Education,* 7(2), 131–153. https://doi.org/10.1007/s40692-019-00150-8

Curriculum Development Council (2001). *Learning to learn: The way forward to curriculum development.* Hong Kong: Government Printer.

Curriculum Development Council (2015). *Ongoing renewal of the school curriculum – focusing, deepening and sustaining: An overview.* Hong Kong: Government Printer.

Curtis, D., & Denton, R. (2003). *The Authentic Performance-Based Assessment of Problem Solving* (Vol. 52). NCVER.

Education Bureau. (2015). *Ongoing renewal of the school curriculum*. Hong Kong SAR Government. https://www.edb.gov.hk/en/curriculum-development/renewal/index.html

Freudenberg, B., Brimble, M., & Cameron, C. (2011). WIL and generic skill development: The development of business students' generic skills through work- integrated learning. *Asia-Pacific Journal of Cooperative Education*, *12*(2), 79–93.

Fitzgerald, G. E., & Koury, K. A. (1996). Empirical advances in technology-assisted instruction for students with mild and moderate disabilities. *Journal of Research on Computing in Education*, *28*(4), 526–553. https://doi.org/10.1080/08886504.1996.10782181

Gibb, J. (2004). Generic skills in vocational education and training: Research readings. *National Centre for Vocational Education Research (NCVER)*.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hui, E. K. (2002). A whole-school approach to guidance: Hong Kong teachers' perceptions. *British Journal of Guidance and Counselling*, *30*(1), 63-80.

Kline, R. B. (2016). *Principles and practice of structural equation modeling, 4th ed*. Guilford Press.

Linacre, J. M. (2018a). *Winsteps® (Version 4.3.1)[Computer Software]*. Beaverton, Oregon: Winsteps.com. https://www.winsteps.com/

Linacre, J. M. (2018b). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon: Winsteps.com. https://www.winsteps.com/

Norman, J. M., Collins, B. C., & Schuster, J. W. (2001). Using an instructional package including video technology to teach self-help skills to elementary students with mental disabilities. *Journal of Special Education Technology*, *16*(3), 5–18. https://doi.org/10.1177/016264340101600301

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2021). *Psych* (2.1.1) [Computer software]. The Personality Project. https://personality-project.org/r/psych/

RStudio Team. (2018). *RStudio: Integrated development environment for R*. RStudio, Inc. http://www.rstudio.com/

Solberg, V. S. H., Lillis, J., Zhang, W., & Martin, J. L. (2020). Career development policy strategies for supporting transition of students with special educational needs and disabilities. In M. Yuen, W. Beamish, & V. S. H. Solberg (Eds.), *Careers for students with special educational needs: Perspectives on development and transitions from the Asia-Pacific region* (pp. 3–16). Springer. https://doi.org/10.1007/978-981-15-4443-9_1

Star, C., & Hammer, S. (2008). Teaching generic skills: Eroding the higher purpose of universities, or an opportunity for renewal? *Oxford Review of Education*, *34*(2), 237–251. https://doi.org/10.1080/03054980701672232

Turnbull, A. P., Turnbull, H. R., & Wehmeyer, M. L. (2004). *Exceptional lives : Special education in today's schools* (4th ed.). Pearson/Merrill/Prentice Hall.

Yang, L., Yuen, M. T., Wang, H., Wang, Z. Y. & Sin, K. F. (2020). Assessing career life skills self-efficacy of students with special educational needs: A comparative study in Hong Kong. In M. Yuen, W. Beamish, and V. S. Solberg (Eds.), *Careers for students with special educational needs:*

Perspectives on development and transitions from the Asia-pacific region (313-326). Singapore: Springer.

Yuen, M., Chan, R. M., Gysbers, N. C., Lau, P. S., Lee, Q., Shea, P. M., ... & Chung, Y. B. (2010). Enhancing life skills development: Chinese adolescents' perceptions. *Pastoral Care in Education*, *28*(4), 295-310.

Wright, B., Linacre, J., Gustafson, J., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

About the authors

Dr. Frank Tian-fang YE is a Research Assistant Professor in the Department of Applied Social Sciences at the Hong Kong Polytechnic University. He previously worked as a post-doctoral fellow at the Centre for Special Educational Needs and Inclusive Education, The Education University of Hong Kong.

Mr. Gao Fengzhan is a PhD student at The Education University of Hong Kong.

Dr. Yang Lan is an assistant professor working at Department of Curriculum and Instruction, The Education University of Hong Kong. She is also serving as co-director of Centre for Special Educational Needs and Inclusive Education.

Dr. Hsu Chia-Ling is a manager of Assessment Technology and Research Division at the Hong Kong Examinations and Assessment Authority. She is also serving as senior fellow of Analytics\Assessment Research Centre at The Education University of Hong Kong.

Prof. Sin Kuen Fung is a Professor of the Department of Special Education and Counselling and Director of Centre for Special Educational Needs and Inclusive Education at The Education University of Hong Kong.