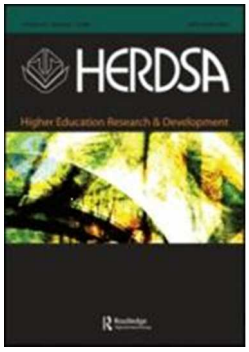

Development of an item bank for assessing generic competence in a higher education institute: a Rasch modelling approach

Journal:	*Higher Education Research & Development*
Manuscript ID:	CHER-2012-0162.R1
Manuscript Type:	Article
Keywords:	Generic skills, Assessment, Statistics, Instrument development, Rasch modelling

Development of an item bank for assessing generic competences in a higher education institute: a Rasch modelling approach

This paper describes the development and validation of an item bank designed for students to assess their own achievements across an undergraduate degree programme in seven generic competences (i.e. problem-solving skills, critical-thinking skills, creative-thinking skills, ethical decision-making skills, effective communication skills, social interaction skills and global perspective). The Rasch modelling approach was adopted for instrument development and validation. A total of 425 items were developed. The content validity of these items was examined via six focus group interviews with target students, and the construct validity was verified against data collected from a large student sample (N = 1151). A matrix design was adopted to assemble the items in 26 test forms, which were distributed at random in each administration session. The results demonstrated that the item bank had high reliability and good construct validity. Cross-sectional comparisons of Years 1 to 4 students revealed patterns of changes over the years. Correlation analyses shed light on the relationships between the constructs. Implications are drawn to inform future efforts to develop the instrument, and suggestions are made regarding ways to use the instrument to enhance the teaching and learning of generic skills.

Keywords: Higher education, generic skills, instrument development, Rasch modelling

Background

Higher education institutions are increasingly being asked to provide evidence of their effectiveness, especially in terms of students' learning outcomes. Students' achievements in generic competences, also known as graduate attributes, 21st-century skills, transferrable skills, and key skills (e.g. Assiter, 1995; Barrie, 2006; Clanchy & Ballard, 1995; Fallows & Steven, 2000), are considered important evidence of the efficacy of higher education institutions. Consequently, the need for universities to provide credible indicators of generic competences has increased rapidly. Universities worldwide have declared the generic qualities they expect their students to develop by the time they graduate. Yet it remains a major challenge to implement these declared aspirations (Dearing Report, 1997). One difficulty lies in the lack of suitable, effective and accessible instruments to assess the set of stated generic skills (Coffield, 1997).

To meet such challenges, universities in the UK, the US and Australia have started to design various instruments to measure graduates' attributes. In the UK, many universities utilise self-assessment tools and learning portfolios to monitor achievements in generic skills (Murphy, 2001). In the US, a nationwide Voluntary System of Accountability Program was implemented in 2010, which recommends three measures of generic skills to over 520 participating higher education institutions (McPherson & Shulenburger, 2006). Among the three measures, the Collegiate Assessment of Academic Proficiency and the Measure of Academic Proficiency and Progress provide test modules for written communication skills and critical thinking while the Collegiate Learning Assessment measures four generic skills: critical thinking, analytic reasoning, problem-solving and written communication skills (Klein, Benjamin, Shavelson, & Bolus, 2007). In Australia, the Australian Council for Education Research has developed the Graduate Skills Assessment to measure four generic skills: critical thinking, problem-solving, written communication and social interaction skills (Hambur, Rowe, & Luc, 2002).

In Hong Kong, the University Grants Committee (UGC), the government funding agency for public higher education institutions, has said the following: 'The quality of teaching and learning should be properly assessed and rewarded on the basis of objective assessment tools and data' (UGC, 2010, p. 5). It further noted that learning outcomes would be used as key performance indicators of the effectiveness of higher education institutions. While all eight UGC-funded universities in Hong Kong have declared their aspirations with respect to graduate qualities, more can be done to develop instruments that assess students' achievements in this regard.

In response to this challenge, a higher education institution in Hong Kong launched a project to develop an instrument to assess students' achievements regarding intended generic attributes. Making references to common frameworks of 21st-century basic skills (e.g. Commonwealth of Australia, 2002; European Communities, 2007; Ministry of Education,

2005) and considering local contexts, the institution identified seven generic intended learning outcomes (GILOs) for its students and graduates based on its own mission statement and strategic targets. These seven GILOs would provide its graduates with a solid foundation for 21st-century work and living in Hong Kong. They are: problem-solving (PS), critical thinking (CT), creative thinking (CIT), effective communication (EC) (oral and written, EC-oral and EC-written), social interaction (SI) skills, ethical decision-making (EDM) and global perspective (GP).

The project team was charged to develop an instrument to measure all seven GILOS, yet we found no existing instrument that assesses all seven GILOs simultaneously in one test set. Most existing instruments assess only one or a few of the seven. Furthermore, most instruments are designed to be rated by teachers, with few for students to rate themselves. Therefore, this project aims to design a reliable, valid and accessible instrument for students to self-assess their achievements in all seven GILOs along their learning trajectory from entry to university to graduation and beyond. As each student will conduct a self-assessment multiple times along his or her learning journey, one single instrument will not be appropriate or reliable for multiple uses. An item-bank approach based on Rasch modelling (Rasch, 1960) was adopted.

This approach is original in four aspects. A Rasch modelling-based item bank is capable of producing multiple parallel test forms, which allow the direct comparison of different students on different occasions, thereby enabling us to track students' abilities and growth on a longitudinal basis (Fischer & Molenaar, 1995; Wright & Stone, 1979). It is also capable of minimising the length of items necessary for reliable estimations of students' abilities. Moreover, the item bank thus developed is open for continuous revision and development. For instance, items may become outdated, and new items can be added. Other important generic attributes that are not yet included in the current item bank, such as information literacy, working in teams and concern for environmental issues, can also be incorporated.

Finally, the item bank could measure seven important GILOs simultaneously. To our best knowledge, no existing instrument has covered as many generic skills within one test set.

This paper focuses on the processes of development and validation of this item bank. It also reports our initial exploration of the growth patterns across student year levels and the relationships between the constructs. The implications for instrument development and use conclude the paper.

Instrument development

The Rasch model-based, measure-construction process (Wilson, 2005) was used to develop and validate the instrument. It contains the following four building blocks:

1. Construct definition: define the constructs with sufficient clarity and specificity for operational uses
2. Item design: develop items that engage the constructs
3. Outcome space: pilot the instrument and assign scores to item responses
4. Measurement model: use the Rasch model to calibrate the items and to establish the psychometric properties of the measures.

Construct definition and item design

Instrument development commenced with the first two building blocks: construct definition and item design. We established a construct map that clearly defined each GILO (i.e. what it was and what it was not). At this stage, an extensive literature review was conducted for each GILO regarding established theories defining its nature and component skills, existing measures and its related variables. Although targeting seven GILOs, we found one GILO, EC, was commonly partitioned into EC-oral and EC-written skills because oral and written communication involve different competences. Therefore, we defined and used eight constructs in the item bank.

On the basis of coherent and sound theory identified in the literature, each construct was defined operationally in terms of its key component skills. For instance, following Curtis and Denton (2003), PS skills are defined operationally as the ability to deal with novel problems, tasks or situations, to plan with existing resources, to execute a plan, to monitor the process and to reflect upon executed solutions. This definition entails four component skills at different stages of PS: to understand a problem, to plan a solution, to execute a plan and monitor the process and to reflect on and evaluate the solutions.

Next, the descriptors for students at different performance levels were written to guide item design. In the case of PS, four descriptors were written to describe performance from beginner to advanced levels. Items were then written for each level in the form of performance tasks or attitude statements. The following are four performance tasks to assess PS skills: (a) be clear about the expected results, (b) identify additional information to help understanding, (c) identify causes, but not symptoms, of the problem and (d) identify ways to improve one's own problem-solving skills. Students rated their own performance on these tasks using a 5-point scale anchored in normative terms as poor (1), below average (2), average (3), above average (4) and outstanding (5). For the two constructs relating to attitudes (i.e. EDM and GP), the students rated their agreement with different attitude statements on a 5-point Likert scale anchored from strongly disagree (1) to strongly agree (5).

In contrast to the traditional approach where item difficulties are specified post hoc based on empirical data collected later, this project designed items on the basis of priori conceptualisations and relevant theories. This entailed hypothesising and describing students' skills at each performance level before collecting data. For this, we raised and attempted to answer questions like: what tasks (items) can students with high/low PS skills perform? Tasks only students with high PS skills can perform are considered to reside at higher performance levels (more challenging), and tasks that both high- and low-ability students can perform are considered to reside at lower levels (easy). Thus the items created are ordered

and categorised in such a way that students who reach a higher level are expected to attain higher scores whereas those who are below a specific level are expected to obtain lower scores. This conceptual measurement model is validated by fitting the hypothetical item difficulty patterns to the empirical data. Verification of the conceptualisation would provide strong evidence to support the construct validity of the instrument.

Content validation

All items were initially written in English and translated into Chinese after completing the preliminary construction process. When the bilingual versions were ready, 27 participants were recruited from the target population for focus group interviews. Participants were assigned to six groups such that each group involved both graduates and current students. Each interview lasted two hours, and two constructs were discussed. Participants first completed a sample test booklet to rate their own performance or attitude. Afterwards, a moderator led participants to discuss the items consecutively, focusing especially on those marked by the participants as problematic. Comments were summarised and addressed. Problematic items were either revised or deleted. The finalised item bank contained 425 items to be validated empirically.

Empirical validation

Item validation consists of the last two building blocks: piloting the instrument and using the Rasch model to calibrate the measures. The 425 items were assembled in parallel test forms and administered to a large student sample. The data was analysed to examine the psychometric properties of each measure and to calibrate the items. Construct validity was established when the 'posterior' response distribution matched the 'priori' hypothetical item structure. Finally, cross-sectional comparisons of Years 1 to 4 students were conducted to explore possible changes to the seven GILOs over time, and multi-dimensional analyses were conducted to explore the relationships between the constructs.

Instrument

Because all items had to be validated simultaneously and it was unrealistic to ask any student to answer all items, the 425 items had to be assembled in short, parallel test forms, and different forms had to be linked via common items. A widely used floating linking item strategy, the balanced incomplete block (BIB) design (van der Linden, Veldkamp, & Carlson, 2004), was used to assemble test forms. The BIB design has the advantage of giving every item an equal chance to serve as a linking item, and the roles of all items are balanced. Twenty-six test forms were assembled, each consisting of about 102 items covering all eight constructs.

Sampling participants

To calibrate the 425 items, we aimed to have around 1000 valid cases in our data. With an expected return rate of 50%, our sampling target was set at 2000 students. Stratified cluster sampling was adopted to select a large sample representative of the target student population. The full sampling framework included all 110 programmes (clusters) of the institute. The strata used in sequence were funding source (self- vs. public-funded), study mode (fulltime vs. part-time), affiliations (all three faculties and the graduate school) and duration (1–4 years). For the first two strata, we selected only UGC-funded programmes that recruit fulltime students. Thereafter, we sampled programmes proportionately from each faculty, covering programmes of all durations. Finally, based on logistical considerations, we only selected large programmes with annual enrolments larger than 75 students. This sampling exercise identified 2466 students from 10 programmes.

Data collection and analysis

Data was collected during programme assemblies held at the beginning of the academic year in September 2010. Altogether 18 programme assemblies were visited. At each assembly, all 26 test forms were distributed in a way that participants sitting next to each other did not

answer the same test thereby minimising method effects and enhancing the validity of individual responses. At the beginning of each session, all of the participants were reassured of anonymity and confidentiality in the handling of the data they provided. Altogether, 1176 students participated but, after deleting invalid cases, 1151 valid cases were kept in the final dataset. This sample size was sufficient for our data analysis.

Rasch modelling with Winsteps software was adopted for the data analysis. For each construct, five major analyses were conducted: (a) evaluate the item dimensionality, (b) examine the item fit to decide if items should be removed from the item bank, (c) assess the spread of item difficulty, (d) compare the hypothetical construct pattern with the empirical data (construct validation) and (e) explore the differences between grades (Years 1–4). Significance tests were conducted to examine the mean differences across the different student groups. The method of plausible values was used to account for uncertainties carried over from the Rasch estimation. Finally, the relationships between the eight constructs were explored using ConQuest software, which analyses multi-dimensional data.

Results

After removing poorly fitted items, the final item bank comprised 342 items with good psychometric qualities (i.e. good item fit and a sufficient spread of item difficulty). Table 1 shows the distribution of items in the item bank.

Item dimensionality

Dimensionality analysis examined the extent that a single dimension could explain the variance of all the items designed to measure a single construct. Eight sets of items were analysed one by one. The results were satisfactory. All eight sets met the uni-dimensionality assumption of Rasch modelling (i.e. each item set could be explained sufficiently by a single dimension), and there was no evidence to indicate a second dimension.

Table 1. Item distribution.

Constructs	Level 1	Level 2	Level 3	Level 4	Sub-total
1. PS	8	14	10	12	44
2. CT	15	13	18	-	46
3. CIT	3	10	26	-	39
4. EC-oral	9	14	20	-	43
5. EC-written	8	14	8	-	30
6. SI	18	17	13	-	48
7. EDM	13	21	12	-	46
8. GP	22	15	9	-	46
Total	96	118	116	12	342

Item reliability and spread of item difficulties

Table 2 demonstrates the reliability indices of the eight item sets. Standardised infit and outfit indices indicated item fitness to the Rasch model. As shown in Table 2, 90% of items had acceptable model fitness with ranges between -2.0 and +2.0. Given the large number of items, this result was considered satisfactory. Moreover, the reliability indices were well above .90 for items and above .80 for persons, which suggested that the sample was large enough to reliably estimate item difficulties, and the items could sufficiently distinguish students with different abilities.

Under the Measure means index in Table 2, we observed the mean values for student abilities were higher than the mean values of item difficulties by at least one logit. This suggests that these students generally considered themselves as proficient in these constructs. However, when such a mismatch pattern is observed, it is important to examine if high-ability students have items that are sufficiently challenging and if low-ability students have items sufficiently easy for their level (Bond & Fox, 2007). To examine the spread of item difficulty, a Wright map of item thresholds was plotted for each item set. Due to space limitations, only the Wright map for problem-solving is presented in Figure 1.

Table 2. Item reliability indices.

Constructs	Statistics	Measure		90% Infit (zstd)		90% Outfit (zstd)		Separation
		Mean	SD	LB	UB	LB	UB	reliability
1. PS	Person	0	1.59	-3.8	2.2	-3.8	2.2	0.82
	Item	-1.28	0.4	-2.3	2.3	-2.6	2.5	0.91
2. CT	Person	0	1.64	-2.5	2.1	-2.5	2.1	0.85
	Item	-1.18	0.36	-2.4	3.3	-2.6	3.3	0.90
3. CIT	Person	0	1.93	-2.9	2.1	-2.9	2.1	0.89
	Item	-1.05	0.46	-2.1	2.1	-2.2	2.3	0.94
4. EC-oral	Person	0	1.65	-3	2.2	-3.1	2.2	0.87
	Item	-1.61	0.47	-3.1	3.3	-2.8	2.3	0.95
5. EC-written	Person	0	1.7	-2.4	1.9	-2.4	2.2	0.83
	Item	-1.04	0.51	-3.4	4.5	-3.5	5.1	0.95
6. SI	Person	0	1.41	-2.4	2.7	-2.5	2.7	0.87
	Item	-1.22	0.1	-2.5	2.7	-2.4	3.0	0.98
7. EDM	Person	0	1.34	-2.3	2.2	-2.3	2.2	0.82
	Item	-1.71	0.1	-2	3.1	-2.2	0.7	0.99
8. GP	Person	0	1.51	-2.8	2.8	-2.9	2.4	0.84
	Item	-2.67	0.92	-4.2	3.7	-3.4	3.8	0.99

This is the pre-published version.

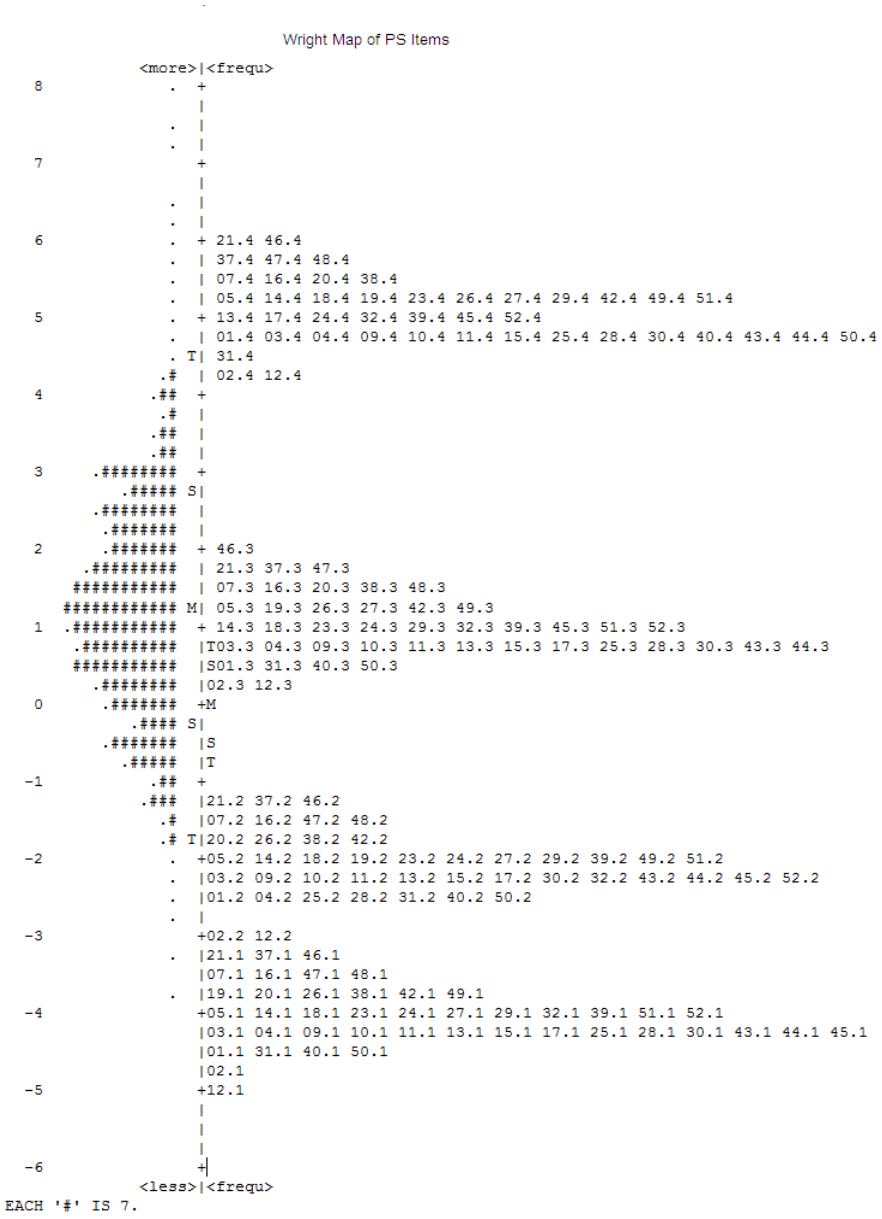

Figure 1. Wright map for the 44 items under problem-solving skills. Students are displayed on the left, and items are on the right. Each # represents seven students in the sample, and the number behind each item indicates specific thresholds on the Likert scale.

The Wright map visually summarises several aspects of the analysis. The distributions of students (on the left) and the items (on the right) are placed on the same logit scale. The numbers behind each item indicate the thresholds of the scale. As each item was rated on a 5-point scale, there were four thresholds for each item, indicating the midpoints (hence

thresholds) between anchor point 1 (poor) and 2 (below average), between 2 and 3 (average), between 3 and 4 (above average) and between 4 and 5 (outstanding). For example, in Figure 1, Item 21.4 refers to the fourth threshold for item 21 (a threshold from 'above average' to 'excellent'), and item 12.1 refers to the first threshold for item 12 (a threshold from 'poor' to 'below average'), these two items being the most difficult (the highest on the scale) and the easiest (the lowest on the scale), respectively. The logit scale has its mean fixed at zero in the middle of the scale. Students located at the same position (or height) as a particular item have a 50% chance of reaching or going beyond the performance level of that item. Students located at a higher position above an item (threshold) have a greater chance of performing better than the level indicated by that item (threshold). Conversely, students located below have a lower chance.

Figure 1 shows the spread of item difficulties could sufficiently cover all student abilities from the lowest to the highest. Most students were located at threshold 3, fewer students were located at thresholds 2 and 4, and few students were located at threshold 1. The items were spread over a large range of difficulty levels (logit values range from -5.0 to +6.0), but the items at the higher end did not seem to be sufficient in distinguishing high-ability students. Specifically, there was a gap between the most difficult item set (threshold 4) and the next item set (threshold 3) and a gap between threshold 3 and threshold 2. Corresponding to these two gaps, there were quite a number of students who had no suitable items at their ability level. To strengthen the differentiation power of the items and improve the item spread, the 5-point rating scale can be changed into a 7-point scale. However, the effect of such a change needs further investigation.

Construct validity

As mentioned earlier, the item bank was designed with performance levels clearly articulated at the beginning of the instrument construction. Each item has a clear prior specification of its

intended performance level. Thus, verification of the model fit between the conceptual patterns and the empirical data provided strong evidence to support the construct validity of the instrument. Figure 2 plots the mean difficulties of the items at each performance level. It shows that the hypothetical structure of the item bank fitted the data satisfactorily. Specifically, items referring to easy tasks showed lower difficulties while items referring to difficult tasks showed higher difficulties. Items designed to be easy (at level 1) were the easiest for the students (with the lowest mean values), and items designed to be more difficult (level 2) were much more difficult than level 1 items. Items designed to be the most challenging (level 3) were the most difficult for the students.

However, for PS and EC-written, the increase from level 2 to 3 was rather flat (mean difference = .029 and .072), suggesting that these two levels were not separated as satisfactorily as the other item sets. For PS, items at levels 2 and 3 could be combined into one level, which would then be distinct from level 4 items. For EC-written, more challenging items at level 3 could be added to the item pool to enhance the spread of item difficulties so that higher-ability students could be better distinguished.

Figure 2. Construct validation: empirical item difficulty mapped against conceptual performance levels.

Further analysis

Two further analyses were conducted: (a) cross-sectional comparisons of Years 1–4 students to explore possible changes over time and (b) correlation analyses to explore the relationships between the eight constructs. Figure 3 presents the results of the cross-sectional comparisons. Table 3 presents the correlation matrix.

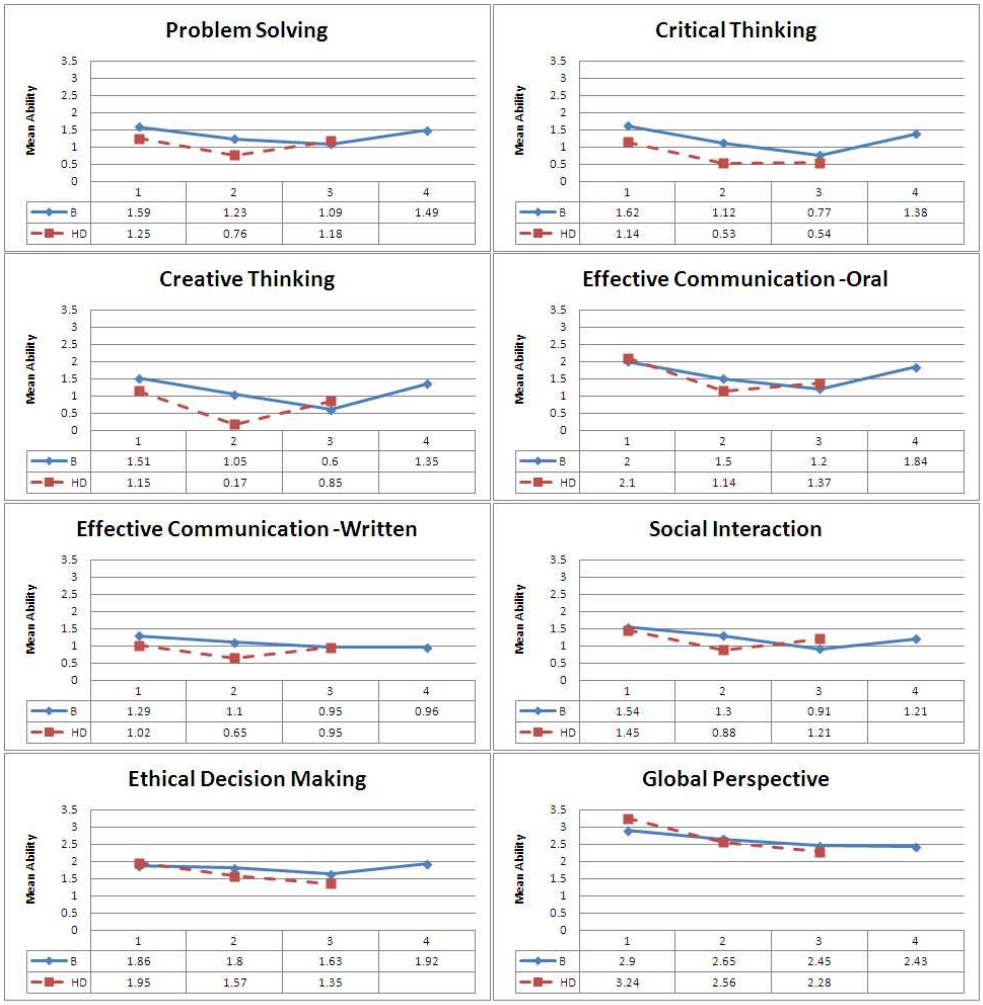

Figure 3. Cross-sectional comparisons of patterns of changes over years.

In Figure 3, the solid line denotes bachelor-degree (B) students where 1 to 4 represent Years 1 to 4 while the dashed line denotes 3-year higher-diploma (HD) students. The HD students were analysed separately from the 4-year B students. As the HD students are those who failed to meet the admission requirements for B programmes, they are expected to demonstrate lower self-evaluation in skills related to academic achievements compared with the B students. Consistent with our expectations, Figure 3 shows the HD group has lower self-rated performance than the B group regarding the four constructs related to academic achievements (PS, CT, CIT, EC-written), especially at Year 1, but not in the constructs (EC-oral, SI, GP and EDM) that are not usually assessed by academic achievement tests.

Furthermore, except for EDM and GP, V-shaped patterns were observed, with the two highest points at the beginning and the end and the lowest point in the middle.

Significance tests were conducted to examine whether the V-shaped patterns were statistically significant or due to random errors. For the B group, we compared the means of two pairs: Years 1 and 3, and Years 3 and 4; for the HD group, we compared the means of Years 1 and 2, and Years 2 and 3. Instead of raw data, student abilities estimated by the Rasch model were used for testing the mean differences. Because the estimated student abilities involved standard errors, directly using ability estimates for mean comparisons would ignore such errors and render inflated significance levels. To achieve more accurate results, error corrections were applied before significance testing according to established procedures (Wu, 2004). The results are reported in Appendix 1.

The results largely confirmed what was observed in Figure 3. Specifically, for the B group, the V-shaped patterns were significant for three constructs: CT, CIT and EC-oral, with both the decline (Years 1–3) and the increase in trends (Years 3–4) statistically significant. The decline in trends for four constructs (PS, EC-written, SI and GP) reached statistical significance, but the increased trends for these constructs were not significant. Finally, for one construct (EDM), neither the decline nor the increase was significant.

Figure 2 and Appendix 1 also suggest that the HD group has similar patterns to the B group with respect to five constructs (PS, CIT, EC-written, SI and GP), but they are different for three constructs: CT, EC-oral and EDM. For CT and EC-oral, the V-shaped patterns could not be established for the HD group in the same way as the B group because the increases from Years 2 to 3 for the HD group were small and not statistically significant. For EDM, the HD group showed a significant decline from Years 1 to 2 (ΔM = .348, p = .011), but Years 2 and 3 remained at a similar level. In contrast, the B group did not show a significant decrease or increase across Years 1 to 4 for EDM.

Table 3 presents the correlation matrix of the eight constructs. The results confirmed our expectations regarding their relationships, lending further support to the validity of our instrument. We expected that the correlations between the three constructs denoting higher-order thinking skills (PS, CT and CIT) would be higher than their correlations with the other five constructs. We also expected that the two constructs relating to students' attitudes (EDM and GP) would be more closely related to each other than to the other six constructs related to skills. Finally, we expected that the relationships between the three constructs relating to communication (EC-written, EC-oral and SI) would be closer than their relationships with the other constructs.

Table 3. Correlation matrix.

	PS	CT	CIT	EC-written	EC-oral	SI	EDM
CT	**.859**						
CIT	**.799**	**.821**					
EC-written	.726	.766	.720				
EC-oral	.613	.649	.655	**.708**			
SI	.531	.510	.637	**.592**	**.873**		
EDM	.384	.380	.412	.391	.441	.467	
GP	.424	.409	.373	.371	.460	.415	**.571**

As shown in the three shaded blocks in Table 3, the correlations between the three thinking skills (from .799 to .859) were higher than their correlations with the other constructs (from .373 to .766). In addition, the two attitude scales were more closely related to each other (r = .571) than to the other six constructs (from .371 to .467). For the three constructs related to communication, the correlations between EC-oral and SI skills (r = .873) were higher than those for any other constructs. This correlation was also the highest in the matrix. There was also a high correlation between EC-written and EC-oral (r = .708), suggesting that writing skills are related to speaking skills (i.e. people who speak well also

tend to write well). However, EC-written was more closely related to the three thinking skills (from .720 to .766) than to EC-oral (r = .708) and SI skills (r = .592). This suggests that the ability to write well is closely related to how well a person thinks. Overall, the results suggest that the instrument achieved good convergent and discriminant validity among the eight constructs. This will be discussed further in the next section.

Discussion and conclusions

This paper has described the process of developing and validating a new instrument for assessing the seven GILOs of one higher education institute in Hong Kong. The final item bank comprises 342 items of good psychometric qualities. The construct validity of the instrument is also well supported. It is however deemed to be desirable to add more difficult items to the bank, especially for CIT skills and EC-written skills, because the performance of students on the items at level 2 is not well separated from level 3. Exploration of the patterns of differences across grades and the interrelationships between the eight constructs produced interesting findings. Three patterns of differences were observed: (a) a V-shaped pattern for three constructs, (b) a successive pattern of decline for four constructs and (c) a flat pattern with no significant decrease or increase for EDM. These patterns, however, may not accurately represent the patterns of change over time for which a longitudinal study is warranted. Due to the nature of cross-sectional comparisons, changes from the entry to the final year, if there are any, may be confounded by cohort differences among student groups. As cohort differences could not be separated from changes over time, we were unable to accurately estimate the extent to which students' generic skills had changed while proceeding with their educational journey. While a longitudinal study is warranted for such a purpose, the patterns observed in this study offer useful insights for further studies.

Firstly, Year 1 students appeared to demonstrate a higher ability than Years 2 and 3 students across the eight constructs. This may be due to cohort effects. The generic skills of

the Year 1 cohort may be better than those of previous cohorts. The introduction of a general education course in secondary schools in Hong Kong may have provided more opportunities for the Year 1 cohort to develop generic skills, such as PS and CT, with the result that they demonstrated a higher level of attainment in these skills. An alternative explanation is that as new entrants to university, first-year students may not fully comprehend the meaning of the GILOs and the sub-skills they entail. Therefore, they may not yet be capable of conducting self-assessments of their generic skills, and their self-evaluations may not be accurate. When making self-evaluations, the first-year students may have rated themselves against their high-school peers whereas the Year 2 and 3 students were more likely to use their more competent university peers as their reference norm. According to social comparison theory (e.g. Corcoran, Crusius, & Mussweiler, 2011; Festinger, 1954), social comparison is central for adults in conducting self-evaluation. Even when given clear reference criteria, adults still tend to evaluate themselves in comparative terms against their own social norms. As the Year 1 students are new university entrants, the social norms against which they evaluate themselves might mainly include those peers who failed to gain admission to university. Therefore, their self-evaluation may tend to be inflated compared with the senior students. 'The campus as a frog pond' metaphor, which says that a frog in a shallow pond feels better than an equally talented frog in a deep pond (Pettigrew, 1967), may apply to this situation.

According to the literature on self-assessments and self-evaluation (e.g., Dunning, Heath, & Suls, 2004; Mabe & West, 1982), self-assessment is a skill that needs to be developed. When comparing students' self-ratings with teachers' marks (see a meta-analysis of 48 studies in Falchikov & Boud, 1989), experienced students (with ≥ three years of enrolment) and graduates tend to provide more accurate self-ratings than less experienced first-year students. However, experienced students also tend to under-estimate themselves. This may explain the decline from Years 2 to 3.

Secondly, this project observed a rising trend in the final year for all six constructs related to skills and competence although only three are statistically significant (i.e. CT, CIT and EC-oral). This may relate to the nature of final-year programmes when students spend a substantial period of time off-campus where they engage with practitioners and professionals in areas such as work placements, internships and practice teaching. Such experience may contribute to their knowledge of their own capacity in handling various concrete tasks, which may correct their tendency to under-estimate themselves and render more accurate estimations. This project found that final-year students had a significant increase in their self-ratings in CT, CIT and EC-oral skills. They also rated themselves higher in PS, SI and EC-written skills although the increases were not statistically significant.

In light of the above findings, follow-up studies may wish to conduct interviews with first-year students to assess their comprehension of the generic skills and to investigate their reference norms in making self-evaluations. For the longitudinal study mentioned earlier, it may be worthwhile to investigate the time-point when first-year students start evaluating themselves against their university peers. This time-point may be a more accurate baseline to start with than the beginning of their university life. Furthermore, the test can be anchored in ways that do not emphasise social norms as such but that are more performance-oriented. For instance, the rating scale can be anchored as poor, satisfactory, good, excellent and exceptional. Furthermore, alternative assessment instruments or perspectives, such as an objective performance test or teachers and peers' ratings, can be employed together with the self-assessment instrument to investigate the extent of convergence and divergence among different perspectives. It would be interesting to see whether or not the disparities between the students' self-assessments, on the one hand, and the teacher and peer assessments and objective performance tests on the other, may narrow down as students mature and grow.

It is interesting to observe that the first-year HD students generally rated themselves lower than the B group in the skills related to academic achievements (PS, CT, CIT and

EC-written). Meanwhile, their self-ratings are similar to the B group in the skills and attitudes not assessed by traditional academic tests (EC-oral, SI, EDM and GP). This is consistent with previous findings on the impact of achievement tests (e.g. Borislow, 1962), according to which achievers tended to have higher self-evaluations of competence than under-achievers. Regarding the difference across grades, although this project observed similar patterns for the HD and B groups, it also observed differences between the two groups (especially in EDM). Such differences may warrant further investigation to ascertain whether they are related to the duration and/or the quality of the two programmes.

Our multi-dimensional analysis provides a preliminary understanding of the relationships between the eight constructs. The correlation matrix largely confirmed our expectations regarding the three major blocks: higher-order thinking skills, communication skills and attitudes. Indeed, their 'internal correlations' were generally higher than their 'external correlations' with the other constructs. Also, it was interesting – but not unexpected – to find that EC-written skills were highly correlated with the three thinking skills and EC-oral skills were highly related to SI skills. A full discussion of the interrelationships between the eight constructs is clearly beyond the scope of this paper. Yet the verification of our heuristic anticipation lends further support to the convergent and discriminant validity of the instrument. That said, like all self-assessment tools, our instrument suffers from the problem of subjective bias, especially for the six constructs denoting skills. As discussed earlier, students' self-evaluation of their skills may not always be accurate.

In closing, we wish to add a few notes about the importance of the generic skills and how universities can – and should – develop these skills. The seven GILOs represent the kinds of skills and attitudes required for a knowledge-based society, which are increasingly demanded by employers and key stakeholders. As such, universities worldwide are encouraged to develop and cultivate these essential qualities. To ensure that teachers and students engage in the teaching and learning of these qualities, universities can make use of

instruments such as the one described in this paper to systematically collect feedback from students, graduates and their employers. Such feedback can be used to identify and monitor areas in need of improvement to inform ongoing programme development. Moreover, GILOs should be embedded within courses and programmes. Ideally, a curriculum-mapping exercise should be conducted to align the courses within a programme with the GILOs so as to ensure that the programme curriculum provides sufficient opportunities for students to develop all essential qualities. Furthermore, universities can make use of their quality assurance mechanisms to monitor teaching, learning and assessment practices within classrooms and to ensure they align with the intended generic skills.

References

Assiter, A. (1995). *Transferable skills in higher education*. London: Kogan Page.

Barrie, S. (2006). Understanding what we mean by the generic attributes of graduates. *Higher Education* (51), 215–241.

Borislow, B. (1962). Self-evaluation and academic achievement. *Journal of Counseling Psychology, 9*(3), 246–254.

Clanchy, J., & Ballard, B. (1995). Generic skills in the context of higher education. *Higher Education Research and Development, 14*(2), 155–166.

Coffield, F. (1997). A tale of three little pigs: Building the learning society with straw. In F. Coffield (Ed.), *A National Strategy for Lifelong Learning* (pp. 44–58). Newcastle, England: University of Newcastle.

Commonwealth of Australia. (2002). *Employability skills for the future*. Canberra, Australia: AusInfo.

Corcoran, K., Crusius, J., & Mussweiler, T. (2011). Social comparison: Motives, standards, and mechanisms. In D. Chadee (Ed.), *Theories in Social Psychology* (pp. 119–139). Oxford, UK: Wiley-Blackwell.

Curtis, D., & Denton, R. (2003). *The Authentic Performance-Based Assessment of Problem-Solving*. Station Arcade, South Australia: National Centre for Vocational Education Research.

Dearing Report. (1997). *Higher Education in the Learning Society*. London: Her Majesty's Stationery Office.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*(3), 69–106.

European Communities. (2007). *The key competences for lifelong learning*. Luxembourg: Office for official publications of the European Communities.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430.

Fallows, S., & Steven, C. (Eds.). (2000). *Integrating key skills in higher education: Employability, transferable skills and learning for life*. London: Kogan Page.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117–140.

Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York, USA: Springer.

Hambur, S., Rowe, K., & Luc, L.T. (2002). *Graduate skills assessment: Stage one validity study*. Camberwell, Victoria: Australian Council for Educational Research, Australia. Dept. of Education, Science, Training.

Klein, S., Benjamin, R., Shavelson, R., & Bolus, R. (2007). The collegiate learning assessment: Facts and fantasies. *Evaluation Review, 31*(5), 415–439.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296

McPherson, P., & Shulenburger, D. (2006). *Improving student learning in higher education through better accountability and assessment*: *A discussion paper*. National Association of State Universities and Land-Grant Colleges (NASULGC).

Ministry of Education. (2005). *Key competencies in tertiary education*. Wellington, New Zealand: Ministry of Education.

Murphy, R. (2001). *A briefing on key skills in higher education*: *Assessment Series No.5*. New York: Learning and Teaching Support Network.

Pettigrew, T. F. (1967). Social evaluation theory: convergences and applications. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (Vol. 15, pp. 241–311). Lincoln: University of Nebraska Press.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

University Grants Committee. (2010). *Aspirations for the Higher Education System in Hong Kong*: *Report of the University Grants Committee*. Hong Kong: University Grants Committee.

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (2004). Optimizing balanced incomplete block designs for educational assessments. *Applied Psychological Measurement, 28*(5), 317-331.

Wilson, M. (2005). *Constructing measures: An item response modeling approach (Vol. 1).*Mahwah, NJ: Lawrence Erlbaum.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

Wu, M. (2004). Plausible values. *Rasch Measurement Transactions, 18*(2), 976–978.

Appendix 1

Table 4. Significance tests of mean differences.

		PS		CT		CIT		EC-Oral	
	Pairs	$\triangle M$*[1]	P*[2]	$\triangle M$	p	$\triangle M$	p	$\triangle M$	p
B	Year 1–3	.449	.029	.887	.000	.965	.000	.765	.000
	Year 3–4	-.314	.167	-.642	.008	-.772	.006	-.624	.019
HD	Year 1–2	.514	.006	.587	.003	.980	.000	.987	.000
	Year 2–3	-.437	.070	-.018	.723	-.684	.016	-.190	.408
		EC–Written		SI		EDM		GP	
	Pairs	$\triangle M$	P	$\triangle M$	p	$\triangle M$	p	$\triangle M$	p
B	Year 1–3	.590	.010	.662	.000	.217	.251	.485	.008
	Year 3–4	-.262	.286	-.265	0.198	-.288	.200	-.021	.716
HD	Year 1–2	.382	.058	.561	.000	.348	.011	.656	.000
	Year 2–3	-.318	.240	-.306	.240	.261	.217	.242	.331

Note:

*1: ΔM = adjusted mean differences

*2: All p values are Sidak corrected; one-tail p values are used due to the prior expectation of the direction of the mean differences. P values above .05 are considered significant.