Title	Using Item Maps in Assessing Learning.
Author(s)	Lee Ong Kim
Source	*Teaching and Learning, 23*(2), 131-144
Published by	Institute of Education (Singapore)

Teaching and Learning, Vol. 23 No. 2 (December 2002) pp. 131–144

Using Item Maps in Assessing Learning

Lee Ong Kim

Abstract

In the monitoring of student progress in the teaching-learning process, it is crucial to be able to identify what students can already do and what areas they still need help in. Scores obtained from tests administered from one time point to the next may not be able to show growth and progress because subsequent scores may remain numerically the same or even "fall" though students have actually acquired extra learning during that period of time. Equating these tests and putting their items onto a common yardstick will enable us to see a distribution of the subject content in their order of increasing or decreasing difficulty levels. This distribution of items on a difficulty yardstick, forms an item map. Students' measures that are calibrated onto the same metric will make it possible for us to directly compare their abilities against the item difficulties. An ability measure that corresponds to a given difficulty level on the common scale would mean that students have higher chances of answering correctly, items below that difficulty level. Similarly, they will have a lower probability of answering correctly, items with difficulty levels above their own ability level and these are the areas where teachers can help their students.

Introduction

It may not be immediately obvious that proper assessment procedures for continuous assessment in schools is an important factor that will give a tremendous boost to teachers' professionalism. Assessment provides feedback on the effectiveness of the teaching, on whether or not the test itself is functioning as intended, and on students' specific misunderstanding of concepts on a given topic or subject that was taught. Timeliness of tests, for example, is one of the important factors that teachers decide upon as a professional. Schools should not require teachers to test students at a given fixed frequency such as "weekly tests" or "monthly tests" as this can cause teachers to lose sight of the fact that assessment serves several important functions in the teaching-learning process that requires a good understanding of when it is most appropriate to test. The teacher is indeed the best person to know when it is best to test the students. While timeliness of testing is an important professional judgement to be made by teachers, it is even more important for teachers to be able to judge student progress or student growth in the subject that is being taught. The purpose of this paper is to highlight

how raw scores mislead teachers on student performance and to suggest the use of "measures" that can be achieved through item calibrations. This paper explains why a positive answer to the question "can a teacher really make conclusions on student growth by looking at the raw scores they obtained from one test to the next?" is not correct.

Raw Scores

Up to this day it is common practice in schools to use raw scores, usually between 0 to 100, as an expression of the status of a student's ability level at a given point in time. In a subsequent test, say a month later, student A may obtain five raw score points more than before while another student, B, may score five points less than his previous score. Under such circumstances, comments by teachers would generally leave an impression that student A has "improved" while student B has "dropped" and should work harder. Has A really improved? Can B retrogress after following an additional month of lessons? Looking at raw scores will not enable us to answer these questions because nothing in these raw scores tells us anything about the difficulty levels of the two tests. Raw scores are sufficient statistics for the determination of person measures and item difficulty calibrations, but by themselves are not measures (Wright, 1992).

Rasch Measures

Rasch analysis calibrates items and measures persons on the same logit scale so that person ability and item difficulty are directly comparable. If a person's measure on a given variable is below the calibration of an item, we can safely conclude that the probability that this person responds correctly to the item, is small. Conversely, a person attempting an item whose difficulty calibration is below his ability level, has a high probability of getting it right. Comparing person measures with item calibrations on a common yardstick, gives us a clear picture of what, in terms of content, a student has mastered, and what concepts may still be unclear to him. The comparison, in other words, is with a criterion measure.

In a mathematics test equating study, 14 test levels from each of two test forms of the Iowa Test of Basic Skills (ITBS) which were used by the Chicago Public Schools (CPS) were equated. These are the Form 7 (levels 7 through 14) and the CPS90 (levels 7 through 14). Each level of the Form 7 and CPS90 tests are parallel forms of the same test. Level 7 of either test form is for Grade 1 students, Level 8 for Grade 2, and so on. The test levels that were equated were therefore two test forms, both with levels meant for Grades 1 through 8. Equating is achieved by linking the response strings of the various test levels through the use of common items (also referred to as overlapping items) and/or the use of common persons (also referred to as overlapping persons), or both. For both Form 7 and CPS90,

levels 7 and 8 do not have overlapping items with each other or with any of the other levels. For these, common persons therefore needed to be used. From level 9 onwards for each test form, there are overlapping items between adjacent pairs of levels and equating can therefore be done through common items.

Calibrating A Single Test Level

Each test level is first calibrated by running a Rasch Analysis using the computer program Winsteps (Linacre & Wright, 2000). The students' responses were first set in the form of a matrix as shown in Fig. 1. Each row of the matrix is for one student consisting of his ID or name, and his responses. If other variables related to the students and schools are required, these can be made to occupy columns before or after the responses to items. In the matrix shown, the students' ID numbers have eight digits and they occupy the first eight columns. This is followed by two empty columns while the responses to Item 1 begin in column 11 for each student. Figure 1 shows part of the actual data for one of the ITBS Mathematics test levels known as CPS 90 Level 7, which is for Grade 1 students. This test level has 82 items and for this study, was taken by 365 students. Figure 1 shows the responses for only the first and last few students, and for the first 70 items only. It would be interesting to point out that this method is sufficiently robust (statistically speaking) to analyse a two-item test taken by two persons. Clearly, however, the standard error in such a situation would be extremely large. The larger the number of items and persons used, the smaller the standard error of measurement (SEM). The Optical Mark Reader (OMR) used to read the data was programmed to return a "letter", that is A, B, C, or D, if the student answered correctly, but to return 1, 2, 3 or 4 if the student chose A, B, C, or D respectively but the response was wrong. Multiple responses was returned as "9" while a non-response was returned as "0". These information are necessary to prepare the control file for the analysis.

In addition to item calibrations (measures of item difficulty levels) and person measures (measures of student ability levels), this analysis also provides us with the test's item separation reliability coefficient. This coefficient tells us how well spread out the item difficulty levels are, to adequately cover the range of person abilities. In addition, misfitting items are identified, thereby enabling the teacher to study why these items are flagged out as misfitting. Common reasons why items are misfitting include ambiguity of the item and guessing by students. These reasons, however, should be investigated, as items can also appear misfitting if they are responded to in other unexpected ways such as "accidental errors" where easy items are answered wrongly by persons with high ability or very difficult items are answered correctly by low ability students. Through the process of item fit analysis, only functioning items will be left for the teacher to use for the measurement of student ability.

```
30531868  BAC2AB1ABC2CABBAA1B2BAA1BA1C3B2C34AD1B442CBA1CDBABDA2B14DCB4CAABDACABA . . . .
30532007  BACAAB231CCCA31AAC1C1A31B31C3311BCADCB3BD10ACCDBABD2A11ADCBBCAABDACA1A . . . .
30532058  BACAAB1A1CCCA3BAACB2B3A2BA1CAB2CBC4DCB2BDCBA4CDB4B322BCADCBB4AA1DA2ABA . . . .
30614801  BA2AA12ABCCCAB1A31BCBAA2BA3C3BCCBCADCBABDCBACCDBABDA2BCADCBBCAABD24ABA . . . .
30658027  BACAA322BC1C331A2C1CB2A23A1C20000CADCB43D4BA2CDB44D3ABCADCBBCAAB1ACABA . . . .
30660064  BA1AA123BCCCA3BA3CBCBAA2BAB1ABCCBCA2CBABDCBACCDBABDA2BCADCBBCAABDA2ABA . . . .
.     .       .       .       .       .       .       .       .       .       . . . .
.     .       .       .       .       .       .       .       .       .       . . . .
.     .       .       .       .       .       .       .       .       .       . . . .
33439997  BACAAB2A3CCCA33AACB23AA2BA3C232C3CADCB3B2CBAC4DBAB33AB4ADCBBCAABDA2ABA . . . .
33440383  BAC3ABCAB1CCA3BAACBCBAA2BABC332CBCADCB4BDCBAC4DB4BD44B4ADCBBC2A3DACABA . . . .
33440472  BACAAB2AB1CCABBAA1BCBAA1BA1CA32CBCADCB31D2BAC2DBABD4AB4ADCBBCAABDACABA . . . .
33440960  1ACAA32ABCCCA31A3CB23AA2B33CA3CCBC33C33B3CB3CC3BA333ABC214142A4BDACABA . . . .
```

Fig. 1. The Data Matrix.

A plot of person ability and item difficulty on the same scale enables us to make direct comparisons between them. An example of this plot is shown in Fig. 2 where each "#" represents 3 persons and each "." represents one person. The figure shows that Question 24 is the most difficult item on this test, followed by Questions 39, 52, 7, and so on. Since the person ability and item difficulty have been calibrated on to the same logit scale, comparisons can be made directly between persons and items, informing teachers which student finds which items easy and which items difficult. This makes it possible for teachers to provide individualized instruction, focusing on each student's main problems or misunderstanding of concepts. Students who are ready to move on to the next step can do so without the hindrance that they will invariably experience when group instruction is used. The slower ones will have more opportunities for closer attention rather than be left behind as would normally happen in group instruction. Teachers who wish to conduct "remedial" classes will be able to identify more accurately each individual student's needs. From Fig. 2 for instance, the teacher will focus on concepts tested by Questions 7, 52, 39 and perhaps even Question 24, for students whose ability measures are around two logits. It will not be wise for the teacher to focus on these same items for students measuring around –1 logit, for instance, as the more immediate concern for these students will be items 13, 32, 34, 37, 59, 68, 36 and so on. This is not to say that all students whose measure is -1 logit must necessarily have difficulties with these items or even items slightly more difficult than these. There will be those who happen to understand certain concepts better than other concepts and will be able to answer questions on them, although the analysis shows that these students would generally find such questions difficult to them. Teachers can determine which items and their concepts are really problematic to whom thereby giving better and more direct focus on specific difficulties faced by particular students.

Table 1 shows the item difficulty calibrations in their measure order for the analysis of CPS90 Level 7. Only the ten most difficult items, the most difficult being Question 24 with an item calibration of 3.20 logits, and the ten easiest items, the easiest item being Question 5 with a difficulty calibration of –2.31 logits, are shown in the table. Table 2 shows the person measures (students' ability measures) in measure order for the top ten and bottom ten students.

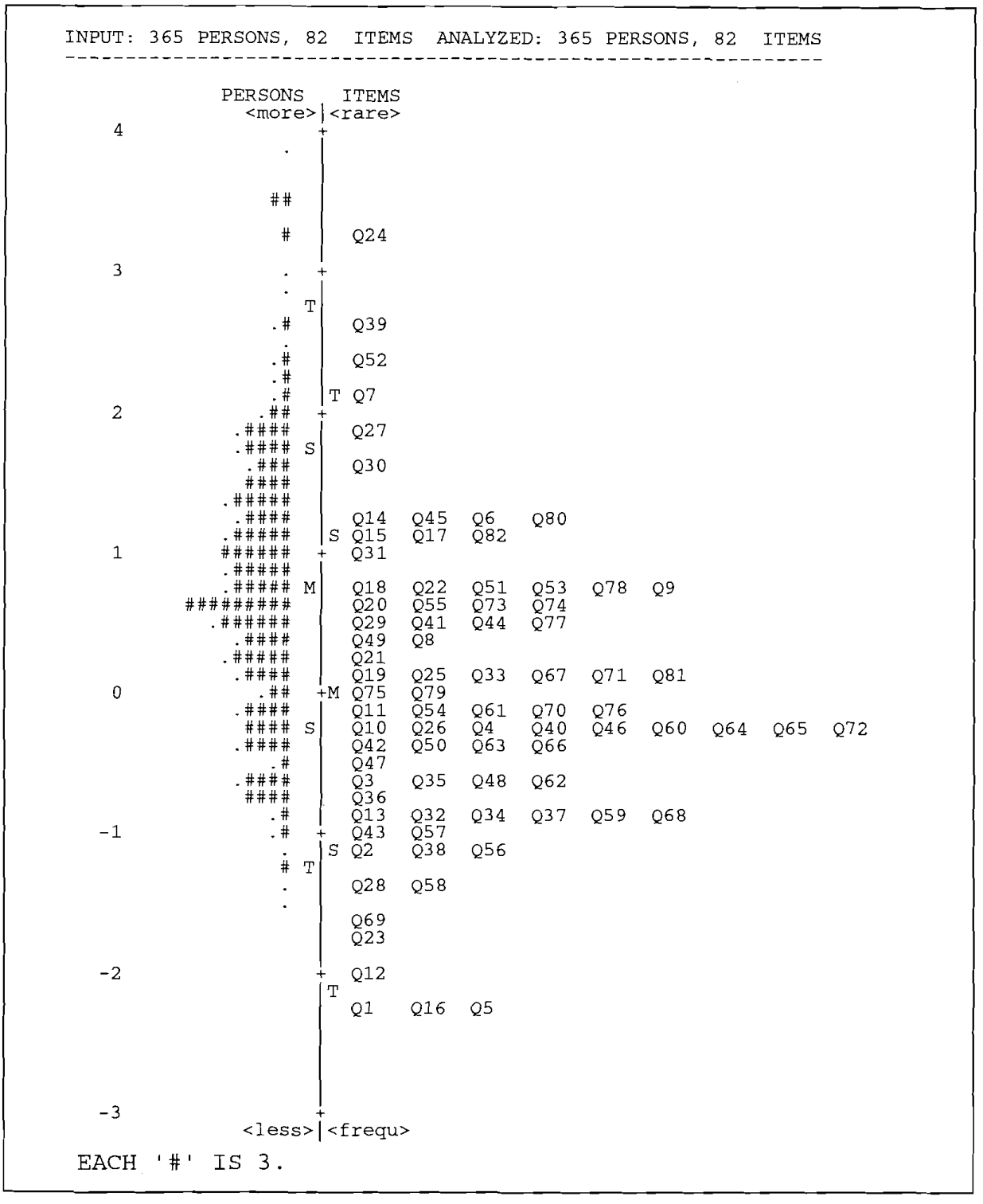

Fig. 2. Map of persons and items.

Calibrating and Equating the ITBS Test Forms and Levels

The ITBS Mathematics consists of multiple-choice items that test on Mathematics Concepts (Math 1), Mathematics Computational Skills (Math 2) and Problem Solving (Math 3). Two forms of the test known as CPS90 and CPS91 follow the same pattern of overlapping items for Math 1, Math 2 and Math 3 between the test

Table 1.
Item Statistics in Measure Order.

Entry Number	Raw Score	Count	Measure	Error	Item Number
24	41	363	3.20	0.18	Q24
39	61	360	2.65	0.15	Q39
52	74	362	2.38	0.14	Q52
7	90	356	2.07	0.13	Q7
27	106	363	1.82	0.13	Q27
30	113	347	1.61	0.13	Q30
6	145	363	1.24	0.12	Q6
45	145	363	1.24	0.12	Q45
80	97	221	1.20	0.15	Q80
14	148	362	1.19	0.12	Q14
.	.	.	.	.	.
.	.	.	.	.	.
56	302	359	−1.18	0.15	Q56
38	302	360	−1.19	0.15	Q38
58	310	362	−1.33	0.16	Q58
28	312	360	−1.42	0.16	Q28
69	318	360	−1.60	0.17	Q69
23	323	362	−1.70	0.18	Q23
12	332	363	−1.98	0.19	Q12
1	337	362	−2.22	0.21	Q1
16	339	362	−2.31	0.22	Q16
5	339	362	−2.31	0.22	Q5
Mean	223	349	0.00	0.14	
S. D.	70	35	1.07	0.02	

INPUT: 365 PERSONS, 82 ITEMS ANALYZED: 365 PERSONS, 82 ITEMS

levels. Level 7 is for Grade 1 students, Level 8 for Grade 2, and so on until Level 14 which is for Grade 8 students. This study takes advantage of the overlapping items for the purpose of equating the test levels. It is essential that the test levels be equated so that a more able student answering higher level tests, will be shown to have higher ability measures even though he obtains the same raw score, say 70%, as students in lower grades.

The pattern of overlapping items from Level 7 through Level 14 for CPS90 and CPS91 is shown in Fig. 3. Levels 7 and 8 of the tests do not have common items. Equating them to any other level of the test would require the use of common persons. In Fig. 3, the overlapping items are represented by the horizontal lines against each test level, with the number of overlapping items shown in the circles. The item numbers for each level is indicated at the ends of each line. For example, Math 1, Level 9 has 28 items numbered 1 through 28 while Level 10 of Math 1 has

Table 2.
Person Statistics in Measure Order.

Entry Number	Raw Score	Count	Measure	Error	Student ID
50	79	82	3.82	0.61	29643059
98	79	82	3.82	0.61	30155440
47	78	82	3.49	0.54	29642907
49	78	82	3.49	0.54	29643032
62	78	82	3.49	0.54	29644063
64	78	82	3.49	0.54	29644209
71	78	82	3.49	0.54	29644810
72	78	82	3.49	0.54	29674566
60	77	82	3.23	0.49	29643997
.	.	.	.	.	.
.	.	.	.	.	.
3	24	78	–0.96	0.27	23943867
247	23	74	–1.00	0.27	31983975
148	21	67	–1.03	0.29	30867122
57	22	70	–1.05	0.28	29643768
295	21	73	–1.15	0.28	32203124
292	22	81	–1.21	0.27	32202764
350	20	73	–1.26	0.29	33036582
324	21	80	–1.28	0.28	32529445
332	19	79	–1.43	0.28	32531113
135	18	74	–1.44	0.29	30748247
Mean	50	78	0.76	0.29	
S. D.	14	4	1.01	0.06	

INPUT: 365 PERSONS, 82 ITEMS ANALYZED: 365 PERSONS, 82 ITEMS

32 overlapping items numbered from 13 through 44. Sixteen items, namely items 13 through 28 are overlapping between these two test levels. The figure shows all the overlapping items for Math 1, Math 2 and Math 3 up to Level 14.

The process of analysis begins with data cleaning. This is necessary because responses to tests may include those from students who were not taking the test seriously and could have "misbehaved" in the test by way of wild guessing, or showing inconsistent responses to items of similar concepts and difficulty levels. In addition, the markings made by students on the OMR forms may be light, giving rise to doubts as to whether they were indeed marked lightly or are the remains after erasure. The first level cleaning was to visually scan the data strings. Those persons with very few items answered, and response strings with clear patterns showing evidence of "wild guessing" such as "CDDABCDDABCDDAB...", were dropped. Strings indicated by the OMR as being light and with embedded omits as shown by the code "0", were also dropped.

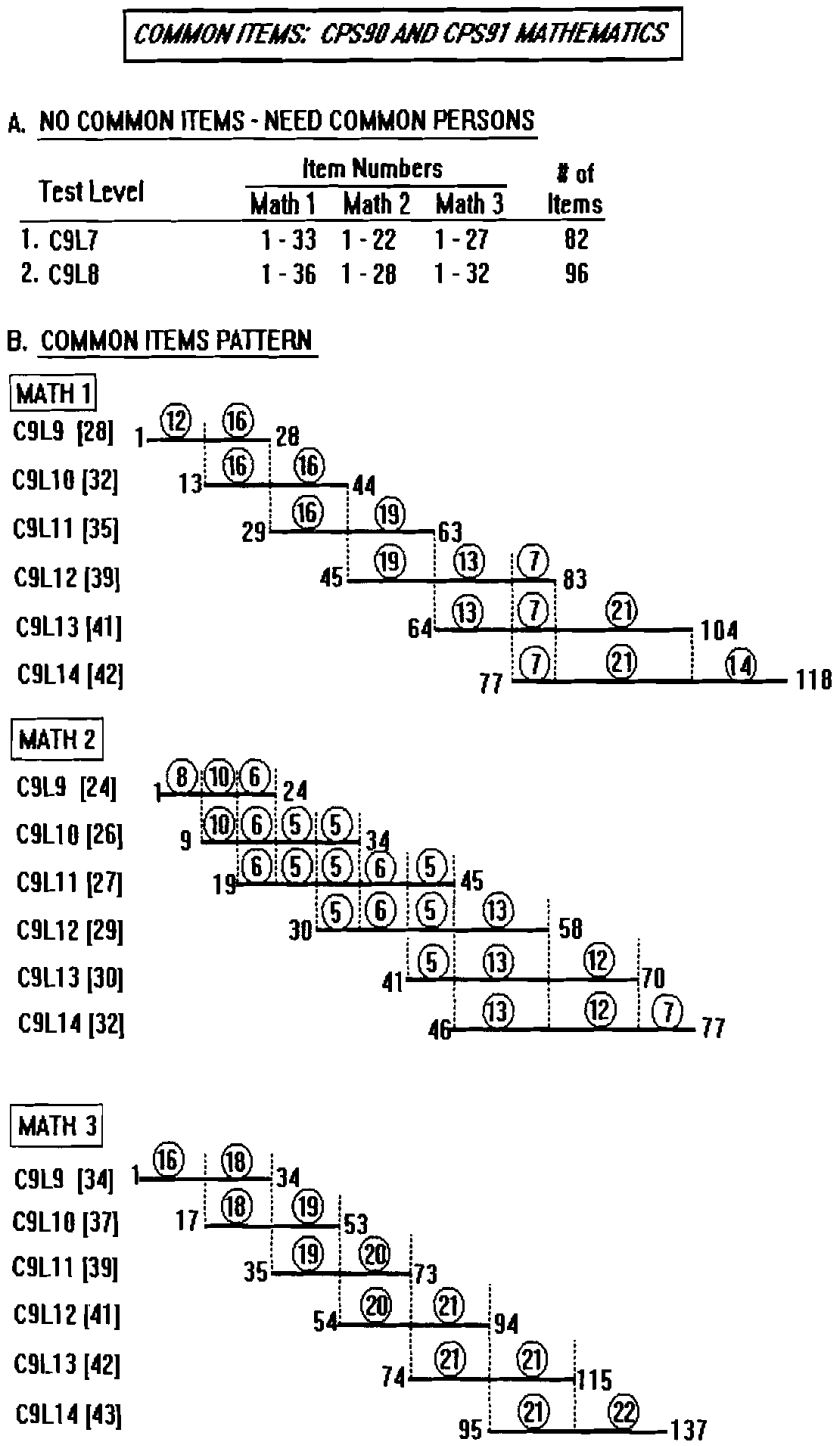

Fig. 3. The overlapping items of CPS90 and CPS91 Mathematics.

Thirdly, persons shown to have both the infit and outfit mean squares to be larger than 2.5 were also dropped from the study because of possible misbehaviours that were not immediately obvious in the test. Lastly, the differences between the performances of common students on adjacent test levels were examined by looking at their standardized differences d_z given by the expression

$$d_z = \frac{[M_1 - M_2]}{\sqrt{(e_1^2 + e_2^2)}},$$

where M_1 and M_2 are each person's measures on the two tests, centred on the mean and e_1 and e_2 are their respective standard errors. The study started with responses from about 30,000 students from which 15% were randomly selected using a programme on SAS. This gave about 4,500 cases. After the cleaning, there were 2,995 persons left in the response matrix, with 1,031 different mathematics items (after taking into account the overlapping items).

From this point there are two ways to equate the test levels. One is to calibrate one test level first, as described earlier, construct an "item anchor file" and then calibrate subsequent test levels with the common items anchored on to the calibrated values from the earlier analyses. In this way, all items of all the test levels involved will be calibrated on to the same scale. Alternatively, the test matrices can be set up with all common persons lined up in the same rows, and all common items lined up in the same columns as shown in Fig. 4 (see Lee, 1992). The horizontal shaded area in the figure shows the shared rows between the matrices of Test 1 and Test 2, representing common persons who took both tests. The vertical shaded area contains the common columns between the matrices of Test 2 and Test 3, representing the common items between the two tests. More test matrices with either common items and/or common persons can be linked in this way to form a single large response matrix. The Rasch analysis is then run on this single matrix.

There is an advantage in running a single analysis on a large matrix compared to running several analyses, one each for each test level by anchoring each of the common items or persons onto the calibrated values of the previous test. The advantage is that the standard error of measurement is very much smaller. Multiple analyses through anchoring can cause the standard errors to add up cumulatively with each run of the analysis.

Calibrated items allow for objective measurement as they can be ordered according to difficulty levels for criterion-referencing of student measures (Wilson, 1992). Ordering calibrated items on a linear scale forms an item map, and one form of this map was shown earlier in Fig. 2. Another way to present an item map to include more information and hence making it easier for teachers to interpret, is as shown in Fig. 5. The mean person ability for each grade was entered into these maps, showing where the students in each grade are likely to be with respect to the content areas. A student located somewhere along this scale, will give an indication to the teacher what he is already able to do and where he still needs help. As an example, some of the criteria from the ITBS are listed under the column titled "Description/Category" in Fig. 5.

The logit scale with its negative values may not be easily comprehensible to users especially those who are comfortable with the "zero to one hundred" scale.

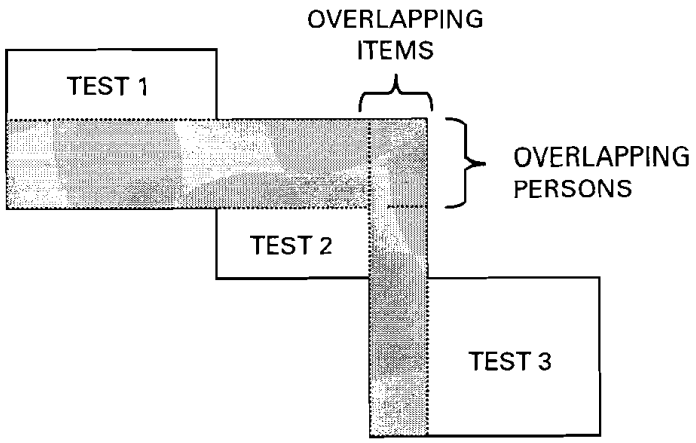

Fig. 4. Matrix for overlapping persons and overlapping items.

We can transform the logit scale into another linear scale, say MC (for Math Concepts), where MC = a + b*(logits). The values of a and b are the "additive" and "spacing" factors (Wright & Stone, 1979, p. 191). The choice of the value for a is such that the origin becomes zero while the value for b is chosen such that rounding off the transformed measures into whole numbers will not cause a lost of interpretive meaning between adjacent measures. The standard error for each calibration is between 0.2 to 0.3. We can therefore usefully distinguish calibration differences of greater than 0.3. Multiplying by 100 is therefore too sensitive and the third digit differences may not indicate any real differences in meanings regarding item difficulties. Multiplying by 10 in this case, seems a reasonable choice. In addition, a translation factor "a" of 55 provides a good range of the new scale MC, from 0 to 105. Both these values of a and b were used in Fig. 5, that is, MC = 55 + 10*(logits). The mean student ability for each grade was positioned in the map. This gives the teacher a guideline as to what content area students of the various grades can be expected to have learned sufficiently well. By comparing his students' abilities with these expected values, a teacher may find it necessary to modify his or her lesson plans in order to assist students who are left behind. Clearly the item map is a very useful tool to use in the interpretation of students' performance.

Conclusion

The concept of student-centred classroom certainly includes the idea of students' self-paced learning. This means that while teachers are assisting slower students to attain the expected learning pace, faster students should not be held back. This invariably happens in a teacher-centred classroom. Identifying misconceptions amongst students require criterion-referenced measures, which raw scores cannot

MC	Logit	Description/Category	Item Example
0	-5.50		
2.5			
		-5.21 Counting objects	C90L7Q1
5.0	-5.00	-5.00 Locating position in line	F7L7Q24
7.5			
10.0	-4.50		
		-4.44 Identifying shapes (circle, etc)	C91L8Q1
12.5			
15.0	-4.00		
17.5			
20.0	-3.50		
22.5			
		-3.17 Numbers in written form (<100)	C90L8Q3
25.0	-3.00	-2.99 Figures representing division	C91L7Q21
		-2.98 Missing number in equation 1 digit)	F7L7Q3
		-2.90 Counting in 'tens' and 'ones'	C91L8Q11
		-2.78 Missing operation in equation	C91L7Q30
27.5			
		-2.72 Concept of 'between'	C91L7Q17
		-2.71 Concept of 'greater than'	F7L7Q19
		-2.55 Identifying coins and summing	F7L8Q28
30.0	-2.50	-2.52 Concept of 'less than'	F7L7Q2
		-2.27 Addition & Subtraction number facts	C91L7Q24
32.5			
		-2.20 Concept of the larger unit	C91L8Q6
		-2.13 Series with negative common diff.	F7L7Q28
		-2.10 Numbers in written form (>100)	C91L9Q2
		-2.07 Identifying number from 'tens' and 'ones'	C90L7Q22
35.0	-2.00		
		-1.89 Series with common diff. > 1	C90L8Q8
	Grade 1		
37.5			
		-1.70 Writing dollars and cents	C90L9Q4
		-1.65 Scale reading (whole numbers)	C90L8Q21
		-1.59 Number in front or behind given position	C90L7Q17

		-1.52 Approximations (up to 2 digits)	F7L8Q26
40.0	-1.50	-1.51 Clock reading	C90L7Q30
		-1.24 Number sentence (> 2 + and/or -); Hundreds	F7L8Q14 F7L9Q18
42.5			
		-1.21 'Closed figure'; Pictorial fractions	C91L9Q24 C91L9Q5
	Grade 2	-1.08 Identifying overlapping regions	F7L8Q22
		-1.01 Mult. & division number facts	C91L8Q26
45.0	-1.00	-1.00 Number sentence (division with 2 numbers only)	C90L10Q29
		-0.97 Number sentence with mult. and division	F7L9Q22
	---	-0.92 Reading given numbers in words	C90L11Q45
		-0.76 Height, length etc. estimates of real objects	C90L8Q2
47.5			
		-0.60 Clock reading with + and - of time	C90L9Q24
50.0	-0.50		
		-0.36 Approximations to products	C91L10Q33
		-0.32 Number sentence (> or = 2 numbers/digits)	C91L8Q28
52.5-		-0.28 Renaming or regrouping tens and ones	F7L8Q32
	Grade 3		
55.0	0.00	0.01 Problems involving larger units	C90L9Q12
57.5		0.25 Multiples	F7L9Q25
		0.36 Concept of odd, even, multiples	C91L11Q48
60.0	0.50		
	Grade 4		
62.5			
		0.90 Scale reading with fractions/ decimals	C91L10Q40
65.0	1.0	0.98 Mission operator (operators on both sides)	C91L8Q34
		1.10 Number sentence mult and division (> 2 numbers)	C91L11Q62
		1.10 Completion of number sentence with > 1 blank	C91L9Q28
		1.19 Rounding to nearest thousand, hundred, ten	C90L12Q67
67.5			
	Grade 5		
70.0	1.5		
		1.67 Rounding decimals to whole numbers	F7L12Q84
72.5			
		1.96 Scale drawing - real distances	C90L12Q70
75.0	2.0		

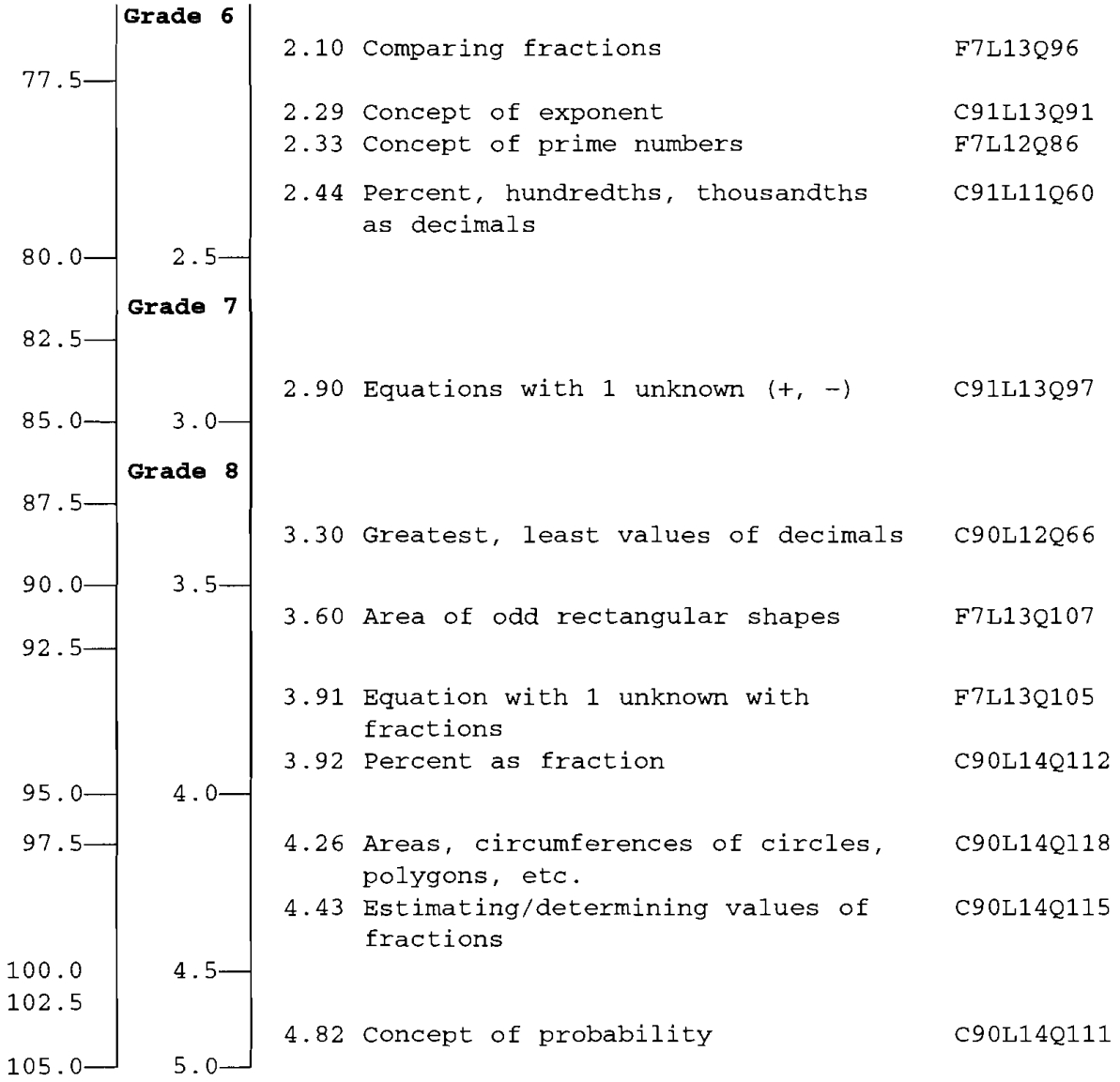

Fig. 5. Item Map for Mathematics Concepts for the ITBS Levels 7-14.

provide. Item maps enable teachers to "zero-in" on the concepts that the different students are having difficulties with. The students can then be given individual attention on the relevant areas for them. Remedial action therefore becomes more effective. With the practice of using Item Maps, it would be heartening for teachers to observe the rapid growth of their students' ability, as they track their performance in their on-going and continual school-based assessment. As new test items are generated by teachers, they can be calibrated and equated to existing items in the same way as described in the test equating section. This approach of coaching individual students shows the vertical integration within the subject area, as there are no artificial curriculum boundaries between the different levels or grades in schools.

Dr. Lee Ong Kim is Associate Professor in the Policy and Management Studies Academic Group at the National Institute of Education. His specialisation and interests include Measurement, Evaluation, and Statistical Analysis (MESA), Rasch Analysis, and Educational Research Methodology

References

Lee, O. K. (1992). Calibration matrices for test equating. *Rasch Measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association.* **6**(1).

Linacre, J. M. & Wright, B. D. (2000). *Winsteps: Rasch Model computer program.* Chicago: MESA Press.

Wilson, M. (1992). Objective measurement: The state of the art, in Wilson, M. (Ed.), *Objective measurement: Theory into practice.* New Jersey: Ablex Publishing.

Wright, B. D. (1992). Scores are not measures. *Rasch Measurement: Transactions of the Rasch Measurement SIG, American Educational Research Association.* **6**(1).

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