DOCUMENT RESUME

ED 343 443 FL 020 294

AUTHOR	Kenyon, Dorry Mann; Stansfield, Charles W.
TITLE	Extending a Scale of Language Proficiency Using Concurrent Calibration and the Rasch Model.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
PUB DATE	Apr 92
NOTE	26p.; Paper presented at the Annual Meeting of the American Educational Research Association (73rd, San Francisco, CA, April 20-24, 1992). For a related document, see FL 020 293.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Chinese; *Comparative Analysis; *Computer Software; *Language Proficiency; Language Research; *Language Tests; Listening Comprehension; Metalinguistics; Models; Reading Comprehension; Second Language Learning; *Test Validity
IDENTIFIERS	BIGSTEPS Computer Program; *Rasch Model

ABSTRACT

The ability of the Rasch model computer program BIGSTEPS to perform concurrent calibration and differences in the outcomes between vertical equating using a common item anchoring method and a common item concurrent calibration method was examined. The data for the investigation came from two Chinese language tests that researchers wished to equate: the extant Chinese Proficiency Test developed in 1984, and a new, lower-level version called the Preliminary Chinese Proficiency Test developed in 1991. Both tests are designed to evaluate the level of general proficiency in Chinese listening and reading comprehension attained by Americans and other English-speaking learners of Chinese. This paper demonstrates the ability of BIGSTEPS to vertically equate a scale of Chinese language proficiency using a concurrent calibration method. In studies on all the items based on the test data, only very minor differences were found between the item difficulty logit values produced by the two methods and those produced when no equating was involved. However, in studies involving only the common items, the concurrent calibration method was found to have a beneficial effect on the calibration of the common items. Users of BIGSTEPS are encouraged to employ the concurrent calibration method in test equating. Contains 4 references. (LB)

Extending a Scale of Language Proficiency Using Concurrent Calibration and the Rasch Model

Dorry Mann Kenyon

and

Charles W. Stansfield

Center for Applied Linguistics

1118 22nd Street, NW

Washington, DC 20037

(202) 429-9494

CAL@GUVAX

Paper presented at the annual meeting

of the American Educational Research Association

San Francisco, CA

April, 1992

One of the practical applications of the Rasch model is to extend a measurement scale. Original methodclogical approaches using common persons or common items were given in Wright & Stone (1979). Since then, available computer programs have simplified the process for users of Rasch measurement. This paper examines two methods to extending a measurement scale using one of the newest Rasch programs, BIGSTEPS (Linacre & Wright, 1991a), in the context of equating two tests of Chinese language proficiency at different levels of difficulty.

The first method examined is an anchoring method. This method involves two separate calibrations. First, one test form is calibrated. The resulting logit values for common items (or common persons) are then used as anchors in the calibration of the second form. The second method examined involves a single, simultaneous calibration of two (or more) test forms linked by common items (or persons). In this paper, we refer to the simultaneous calibration of two tests as the "concurrent calibration" method. This method of equating is discussed in Hambleton & Swaminathan (1985), though available at that time only for the mainframe LOGIST computer program.

At the time the current project began (1990), language testing staff at the Center for Applied Linguistics (CAL) were beginning to employ Rasch methodology in their test development projects. Although the apparent simplicity of the concurrent calibration method was appealing to us, our review of the literature on equating did not indicate that the Rasch program BIGSTEPS had the capacity to perform it. However, the user's manual (Linacre & Wright, 1991b) indicated that the program had the ability to treat items not answered as unreached. Since this capacity enabled LOGIST to perform concurrent calibration, we inferred that BIGSTEPS should be able to do it as well. The literature review also did not reveal any information on what differences there may be in outcomes between the anchoring and concurrent calibration methods. However, we suspected that there may be some advantages, in addition to simplicity, to the concurrent calibration method since concurrent calibration is performed using a fuller set of data for the common items than when those items are calibrated separately, as in the anchoring method.

In this paper, we examine both the ability of BIGSTEPS to perform concurrent calibration and differences in the outcomes between vertical equating using a common item anchoring method and a common item concurrent calibration method. The data for the investigation comes from two Chinese language tests that we wished to equate.

Background to the Tests

The tests to be equated were the extant Chinese Proficiency Test (CPT), originally developed in 1984 by the CAL, and a new, lower-level version, called the Preliminary Chinese Proficiency Test (Pre-CPT), developed in 1991. Both tests are designed to evaluate the level of general proficiency in Chinese listening and reading comprehension attained by Americans and other English-speaking learners of Chinese. The specifications for both tests, especially the Pre-CPT, relate as closely as possible to the scales of reading and listening proficiency established by the Federal Interagency Language Roundtable (FILR) and the American Council on the Teaching of Foreign Languages (ACTFL). On the older CPT, some of the listening and reading stimuli were authentic, in that they were taken from real-life language use. Other stimuli were written by item writers to represent authentic language use. All the listening and reading stimuli on the Pre-CPT are authentic in that they are excerpts from real-life language usage.

For the CPT, stimuli were targeted to the levels Intermediate, Advanced and Superior on the ACTFL proficiency scale. Of the total of 150 4-option multiple-choice items, 60 items test Listening Comprehension, 35 items test Structure and 55 test Reading. In the CPT operational program, examinees are provided with raw scores for the three individual parts, as well a total score. Norms are published by CAL to help score users interpret examinee scores. By 1990, the test had been administered to more than 2000 examinees at than 110 different institutions of higher education.

The CPT was made operational in 1987, and has since been well received. It is fair to say that the CPT has become the standard by which individuals and programs are judged in the Chinese language teaching community. However, in recent years CAL has received many requests for a lower-level version of the test. Accordingly, the Pre-CPT was developed to assess students enrolled in lower-level college and advanced high school Chinese language courses. Items were written targeting the two lowest levels on the ACTFL scale (Novice and Intermediate). The Pre-CPT was field tested in the spring of 1991 on 299 students. Using both Rasch and classical analyses, malfunctioning items were identified and deleted.

The data discussed in this paper was collected during the norming administration of the Pre-CPT. This administration had two objectives. The first was to vertically equate the Pre-CPT with the existing CPT so that scores on both tests could be interpreted on a common scale, thus extending the original CPT scale. The second was to provide preliminary national norms to be used in interpreting test scores.

Method

Subjects In the late spring of 1991, 651 students from around the United States participated in the norming administration of the Pre-CPT. Approximately 49% of the participants were male while 51% were female. Students enrolled in colleges comprised 48% of the total norming sample, students in high schools comprised 50%, and students in weekend Chinese language schools comprised about 2%. Among the sample, 69% were ethnic Chinese. Of the total sample, 15% spoke Mandarin Chinese at home, 32% spoke a language of China other than Mandarin, and a slight majority (53%) did not speak any Chinese at home. Although the proportion of ethnic Chinese in this sample may seem high, it is in fact quite typical, since in many parts of the United States ethnic Chinese make up a large proportion of beginning Chinese language students.

Design for Linking Tests We chose common item equating as the most practical method for this situation. Items from the extant CPT appropriate for inclusion on the norming administration version of the Pre-CPT were identified as follows. First, from the existing CPT database, we conducted a classical item analysis on the responses of the 174 beginning level CPT examinees (i.e., examinees who indicated that they were completing first year college level Chinese). Items that showed appropriate item difficulty and point bi-serial discrimination values for this group were identified. We then examined the performance of the total CPT population from the database on these items. Although these items were generally easy for the entire population, we chose those which retained appropriate discrimination values for inclusion as common items on the norming administration version of the Pre-CPT. Ten Listening Comprehension and ten Reading Comprehension items were selected in this manner. These were interspersed throughout these two sections of the norming administration version of the Pre-CPT.

For the Structure section, finding anchor items was not as straight-forward. The CPT has 35 structure items utilizing two separate item types. The first asks examinees to indicate where within a Chinese sentence a certain Chinese character would be correctly placed. Four possible locations are indicated. The second item type is a single-sentence MC cloze that asks examinees to complete the missing portion of a sentence with one of four options. However, in the Pre-CPT, only one item type, a standard MC cloze, is used in the Structure section. Here, examinees are presented with a paragraph with five words missing. For each missing word examinees are asked to choose the best completion from among four options. The items on the CPT most similar to these were the 20 single-sentence cloze items. Unfortunately, most of these were very difficult for both the beginning-level CPT examinees and the entire CPT population. Of the 20 single sentence items, only six appeared potentially appropriate for the Pre-CPT target group population. Thus, only these six items were used to link the Structure sections of the two tests. On the norming version of the Pre-CPT, these six items were separately presented to examinees as the first part of the structure section of the test. The second part contained the 25 paragraph-level items developed for the Pre-CPT.

Procedures Data from the CPT was compiled as follows. First, the CPT data bank was updated to include all examinees who had taken the test as of June, 1991. Since some examinees take the CPT more than one time, the database used for the analysis was a subset of the complete database, in order that each examinee would appear in the calibration sample only once. Also, not all CPT examinees took every subtest. Thus, for this equating study, the following numbers of CPT examinees were used: 1697 for Reading, 1736 for Structure, and 1697 for Listening. All 651 examinees who took the Pre-CPT norming version were included in this study.

BIGSTEPS was used to calibrate the exams in a sequence of runs. First, each section (Listening, Reading and Structure) on each test (CPT and Pre-CPT) was calibrated separately. This separate calibration was used to establish "reference" values for comparing differences in the two equating methods. Next, the results for the common items in each section of the CPT were used as anchors in a second calibration of the Pre-CPT (Anchoring 1). The item difficulty logits from these two calibrations were combined to form a single scale. Next, the concurrent calibration method was applied. The two data sets were combined, so that responses to each common item formed a single column. For the rest of the columns, items unique to the CPT contained blanks for the Pre-CPT examinees, and items unique to the Pre-CPT contained blanks for the CPT examinees. The results of this calibration (Concurrent) were correlated to the results from the anchoring method. Finally, to further compare the two methods, the anchoring method was again applied. This time, the results of the calibration for the Pre-CPT were used to anchor the calibration of the CPT (Anchoring 2). The item difficulty logits from the two separate calibrations were then combined to form a single scale.

Results

First, results of a classical item analysis for the norming administration of the Pre-CPT are presented in Table 1.

Table 1
Test Statistics from the Pre-CPT Norming Administration

Section	Number of Examinees	Number of Items	Mean Score	Std. Dev.	Reliability	Mean P-value	Std.Dev. P-value
Listening	648	65	52.41	11.12	.94	.81	.11
Reading	642	60	44.51	10.46	.92	.74	.13
Structure	635	31	21.06	6.87	.89	.68	.12

These results indicate that the subtest reliabilities were very high. While the mean p-value for the structure and reading items is appropriate for a multiple-choice test, the Listening Comprehension section may have been a little too easy for this group. This may be due to the presence of native speakers of Chinese in the sample.[1]

An important consideration before equating is to examine the correlation of item difficulty values when the common items are calibrated separately on each test in each population. Table 2 shows the correlations between the common items separately calibrated.

Table 2
Correlations of the Difficulty Values of the Common Items
Calibrated Separated for the Pre-CPT and the CPT Populations

Section	Number of Common Items	Correlation
Listening	10	.91
Reading	10	.92
Structure	6	.49

These figures indicate that the vertical equating of the Listening and Reading Comprehension sections could be performed with confidence. The correlation between the common items on the Structure section was more tenuous. Reasons for the low correlation appear to lie in the fact that there are only six items and all were extremely difficult for the Pre-CPT population. Upon examination of the plot of the calibrations for the structure items, it was noted that two of the six items were much more difficult for the Pre-CPT population than would have been expected.[2] When these two items were dropped, the correlation between the four remaining items was .92. We note here that the actual equating for operational purposes was performed using only four common items; however, in the discussion that follows, all six items are included. Thus, we may note any effects due to the difference between the high correlation on the separate calibrations for the Listening and Reading common items and the lower correlation for the Structure items.

Table 3 presents the correlation between the item difficulty values obtained from the first anchoring method (CPT calibrations used as anchors) and the concurrent calibration for the total number of items equated.

Table 3
Correlations of Difficulty Values in Logits
Between the Anchoring Method
and Concurrent Calibration Method of Equating

Section	Number of Items	Correlation
Listening	115	.9996
Reading	115	.9981
Structure	60	.9984

Table 3 indicates almost a perfect correlation between the logit values using either method of calibration. This provided evidence to CAL staff that BIGSTEPS could indeed handle concurrent calibration.

Table 4 presents the correlations between the item difficulty values in logits from the separate calibration of each test section with item difficulty values resulting from using the concurrent and the anchoring method, respectively. In this analysis, the results of the separate (pre-equating) calibrations serve as a reference against which to compare the post-equating results. Results of both anchoring procedures (CPT used to anchor the Pre-CPT, Pre-CPT used to anchor the CPT) are used as appropriate.

Table 4
Correlations of Difficulty Values in Logits
between Separate Calibrations of Each Test and
Equating Calibrations, by Method

Section/Test (# of items)	Concurrent	Anchor
Listening		
CPT (60 items)	.9991	.9877
Pre-CPT (65 items)	.9891	.9800
Reading		
CPT (55 items)	.9988	.9915
Pre-CPT (60 items)	.9939	.9913
Structure		
CPT (35 items)	.9987	.9846
Pre-CPT (31 items)	.9802	.9690

Although all the correlations in Table 4 are extremely high, there are some small but consistent differences. Correlations between the item difficulty logits calibrated separately were always higher with the concurrent method than with the anchoring method. Correlations for the CPT under either method were higher than for the Pre-CPT. Correlations for the CPT under concurrent calibration were nearly perfect. The lowest correlation was using the anchor method with the Structure section of the Pre-CPT, but even this correlation is not very low.

Table 5 compares the differences in the results from the two methods in terms of absolute values of differences between the logit values when calibrated separately and when calibrated for equating purposes using either the concurrent method (con) or the anchoring method (anc). Table 5 presents the number of items under comparison, the minimum difference between the two logit values for any item, the maximum difference between the two logit values for any item, the mean difference between the two logit values for all items, and the standard deviation of the difference scores. The closer the standard deviation is to 0, the closer the item difficulty logits are to each other in absolute terms.

Table 5
Summary of Values of Difference Scores
Between the Pre-equated Calibrations
and the Equated Calibrations
By Section, Test and Equating Method
for All Items

Listening	N	Min	Max	Mean	Std Dev
CPT-CON	60	-1.36	-.95	-1.14	.05
CPT-ANC	60	-3.14	-1.59	-2.30	.21
PCPT-CON	65	.64	1.79	1.08	.15
PCPT-ANC	65	1.59	3.14	2.29	.20

Reading	N	Min	Max	Mean	Std Dev
CPT-CON	55	-1.22	-.82	-.90	.05
CPT-ANC	55	-2.72	-1.44	-1.96	.13
PCPT-CON	60	.62	1.50	.84	11
PCPT-ANC	60	1.44	2.72	1.96	.13

Structure	N	Min	Max	Mean	Std Dev
CPT-CON	35	-1 02	-.69	-.78	.05
CPT-ANC	35	-2.56	-1.32	-1.86	.18
PCPT-CON	31	.63	1.54	.90	.15
PCPT-ANC	31	1.32	2.56	1.86	.19

Table 5 shows that though the standard deviation of the difference scores are very small, on every test and for every form the standard deviation for the concurrent calibration method is smaller than for the anchoring method. This suggests that this method produces item difficulty logit values after equating that more closely parallel item difficulty logit values before equating. In other words, the concurrent calibration method effects the logit values less than the anchoring method.

In order to examine more closely how the results of the two methods may differ, the above analysis was repeated on the items that were unique to each section. Table 6 presents the results of that analysis.

Table 6
Summary of Values of Difference Scores
Between the Pre-equated Calibrations
and the Equated Calibrations
By Section, Test and Equating Method
for Unique Items Only

Listening	N	Min	Max	Mean	Std Dev
CPT-CON	50	-1.14	-1.13	-1.13	.00
CPT-ANC	50	-2.31	-2.27	-2.30	.01
PCPT-CON	55	1.05	1.08	1.07	.01
PCPT-ANC	55	2.27	2.30	2.28	.01
Reading	**N**	**Min**	**Max**	**Mean**	**Std Dev**
CPT-CON	45	-.89	-.88	-.89	.00
CPT-ANC	45	-1.97	-1.96	-1.96	.00
PCPT-CON	50	.79	.82	.81	.01
PCPT-ANC	50	1.96	1.97	1.96	.00
Structure	**N**	**Min**	**Max**	**Mean**	**Std Dev**
CPT-CON	29	-.78	-.77	-.77	.00
CPT-ANC	29	-1.87	-1.86	-1.86	.00
PCPT-CON	25	.84	.87	.86	.01
PCPT-ANC	25	1.86	1.87	1.86	.00

Table 6 indicates that, as far as the unique items are concerned, there was no difference between the methods of equating. In fact, the item difficulty values of the logits for the unique items appear to be relatively unaffected at all by equating. It may also be noted that the results of a correlational analysis of the unique items reveal correlations above .998 for both methods on all sections of both tests.

Table 7 presents the same analysis on the common items.

Table 7
Summary of Values of Difference Scores
Between the Pre-equated Calibrations
and the Equated Calibrations
By Section, Test and Equating Method
for Common Items Only

Listening	N	Min	Max	Mean	Std Dev
CPT-CON	10	-1.36	-.95	-1.16	.13
CPT-ANC	10	-3.14	-1.59	-2.31	.52
PCPT-CON	10	.64	1.79	1.15	.39
PCPT-ANC	10	1.59	3.14	2.31	.52

Reading	N	Min	Max	Mean	Std Dev
CPT-CON	10	-1.22	-.82	-.97	.11
CPT-ANC	10	-2.72	-1.44	-1.96	.33
PCPT-CON	10	.62	1.50	.99	.22
PCPT-ANC	10	1.44	2.72	1.96	.33

Structure	N	Min	Max	Mean	Std Dev
CPT-CON	6	-1.02	-.69	-.83	.12
CPT-ANC	6	-2.56	-1.32	-1.86	.46
PCPT-CON	6	.63	1.54	1.03	.34
PCPT-ANC	6	1.32	2 56	1.86	.46

Table 7 reveals that it is the calibration of the common items that are affected by the method of equating. In this data set, the concurrent calibration method reduces the standard deviation scores of the difference scores considerably across the tests and sections. This means that the common items are more effected by the anchoring method than by the concurrent calibration method. Table 7 also shows that the standard deviation of the difference score between the test calibrated separately and the concurrent calibration is lower for the CPT than for the Pre-CPT. This may be due to the larger sample size involved in the CPT calibrations which produced more stable item difficulty estimates than for the Pre-CPT.

Table 8 presents a correlational analysis of the data discussed in Table 7.

Table 8
Correlations of Difficulty Values in Logits between Separate Calibrations of Each Test and Equating Calibrations, by Method, For the Common Items

Section/Test (# of items)	Concurrent	Anchor
Listening		
CPT (10 items)	.9900	.9096
Pre-CPT (10 items)	.9586	.9096
Reading		
CPT (10 items)	.9923	.9227
Pre-CPT (10 items)	.9633	.9227
Structure		
CPT (6 items)	.9653	.4905
Pre-CPT (6 items)	.7009	.4905

In Table 8, we see the effect of concurrent calibration quite clearly. The correlation for the anchoring method is, of course, the correlation between the two separate calibrations presented in Table 2. In each case, under concurrent calibration, the correlation between the item difficulty logit values for the equated common items and those from the original separate calibration has risen impressively, particularly in the case of the Structure section. Taken together, Tables 7 and 8 clearly indicate that the concurrent calibration method provides the lesser amount of deviation from an pre-equated solution.

Next, we looked to see if there would be any effect on the item fit statistics due to concurrent calibration. Given the large and disparate sample sizes for the CPT and the Pre-CPT populations, the criterion used was the infit and outfit mean squared statistic provided in the BIGSTEPS output, which is not sensitive to sample size. An item was considered misfitting if both of the mean square fit statistics were greater than 1.20 or less than .80. There are three sets of items to be considered: those unique to the Pre-CPT, those unique to the CPT, and the common anchor items. Table 9 indicates the number of the anchor items that were misfitting under the separate Pre-CPT and CPT calibrations, and misfitting under the concurrent calibration.

Table 9
Number of Misfitting Anchor Items Under Separate and Concurrent Calibrations

	Pre-CPT Separate Calibration	CPT Separate Calibration	Concurrent Calibration
Listening (10 items)			
>1.20	1	0	0
<.80	1	0	1
Reading (10 items)			
>1.20	1	0	0
<.80	0	0	0
Structure (6 items)			
>1.20	1	0	0
<.80	0	0	0

Table 9 indicates that fit was not a problem for the anchor items. None of the anchor items (all of which came from the CPT) in any section were misfitting on the CPT, though two of the anchor items were misfitting in the Listening section of the Pre-CPT, and one was misfitting in the Reading and Structure sections. However, when these anchor items were concurrently calibrated using the entire sample, only one (in the Listening section) remained misfitting.

Table 10 shows the number of the items unique to each test misfitting under separate and concurrent calibration.

Table 10
Number of Misfitting Items Under Separate and Concurrent Calibrations

	Separate Calibration	Concurrent Calibration
Unique Pre-CPT Items		
List (55 items)		
>1.20	4	4
<.80	2	2
Read (50 items)		
>1.20	2	2
<.80	0	0
Str (25 items)		
>1.20	2	2
<.80	0	0
Unique CPT Items		
List (50 items)		
>1.20	4	4
<.80	3	3
Read (45 items)		
>1.20	4	4
<.80	1	1
Str (29 items)		
>1.20	3	3
<.80	0	0

Table 10 indicates that number of misfitting items under separate and concurrent calibration was exactly the same. Upon closer analysis, all of the items misfitting under each calibration were exactly the same, and if their mean square INFIT and OUTFIT ~~statistics~~/differed at all, it was by a maximum of only .01 logits. Tables 9 and 10 indicate again that concurrent calibration effects only the common items and not the unique items on the tests to be equated.

Discussion of the Results

The results above demonstrate the ability of BIGSTEPS to equate tests using both a concurrent calibration method or an anchoring method. In an effort to compare the two methods, the results of the separate calibration of each test section for each test were used as the "reference" against which to compare the results of the two different methods of equating.

The analyses above suggest that only the calibration of the common items are actually affected by either equating method. The relationship of the pre-equated and post-equated item difficulty logit values was unaffected by either method of equating.

In addition to its simplicity, the concurrent method appears to be preferable when the results of the separate calibration are being used as a reference. The results of this study indicate that the concurrent calibration method increases the correlation between the post-equated item difficulty values for the common items and their original pre-equated values, and decreases the variance in the differences between their respective absolute values.

However, that only the common items are affected by the equating method suggests that any analysis of the differences between the two methods would be influenced by the effect of the common items on the entire set of items to be equated. The more common items present, the greater the potential influence. With many common items, to the degree that there is a low correlation between the common items when separately calibrated, the comparison of the anchor method to the concurrent calibration method will be correspondingly poorer. With few items, as in this study, or a higher correlation (which should generally be the rule), differences between the two methods might be minimal.

In terms of fit statistics, the concurrent calibration was able to influence the post-equated fit of common items when compared to pre-equated fit. This result, however, may be heavily influenced by the size of the samples used in equating. In the current example, almost three times as many examinees were present for the CPT than for the Pre-CPT. In three out of four cases, common items that were misfitting in the separate Pre-CPT calibration did not show misfit when calibrated concurrently with the entire sample. The only case where this failed to happen was with an common item that was close to misfitting for the CPT population as well. Had the CPT sample size been smaller than that of the Pre-CPT, the fit statistics may not have been so heavily influenced.

The low correlation between the common items in the Structure section when separately calibrated appeared not to negatively influence overall outcomes, as in Tables 4 and 5. In a sense, this is troubling. For the purposes of this study, equating proceeded mathematically for the Structure section of the two tests, although there may not have been enough evidence to support it. This is again a reminder that theory must drive numbers.

Outcome of the Equating

Encouraged by the results reported above, we used the concurrent calibration method to vertically equate the new lower-level Pre-CPT to the older higher-level CPT. However, since only raw scores had been used in the CPT program before, we needed to convert logit scores into scale scores. Since CAL's Chinese language testing program had always been interpreted in terms of norms, we used normative scaling units, with a mean performance of all examinees (Pre-CPT and CPT) as 100 and 20 as a standard deviation. For the Pre-CPT, scores were based on the performance of the examinees on the items retained in the final form of the test, rather than on the total number of items in the norming administration version. The final form excluded five of the ten common items from the Listening and Reading sections, and all six of the common items from the Structure section. In addition, to shorten the test, ten additional items were removed from the Listening section and five from the Reading section of the Pre-CPT, leaving 50 items in each section.

In order to establish a preliminary table of norms for the Pre-CPT, we estimated scale scores for the final version of the Pre-CPT for all the Pre-CPT examinees who participated in the norming administration. To do so, we used the item calibrations from the concurrent calibrations as anchors. The person ability estimates in logits were then converted to the scale score using the appropriate equations. Ability estimates for persons with perfect scores were calculated with the default value of BIGSTEPS. As noted above, only four of the six common items were actually treated as common items in the final equating of the Structure section. The two other common items were treated as unique items on both tests.

Scale scores for the CPT were similarly determined by a separate calibration of all examinees in the CPT database, not just those used in the concurrent calibration. Thus, if an examinee took the CPT more than once, he or she received an ability estimate for each occasion the test was taken.

In order to assess how well the goal of building a single scale for the two tests was accomplished, we developed a table of scale score means for each examinee subgroup for each section for each test. These are presented in Table 11. For the Pre-CPT, subgroups are divided by year in high school and whether the examinees speak Mandarin, another Chinese language, or English at home. For the CPT, subgroups are divided by college course level and whether Chinese or English is indicated as the native language. In Table 11, the means for each section are given on the first line in bold. Underneath each mean is its standard deviation. On the bottom line, in parentheses, is the number of examinees in the subgroup. For the Pre-CPT, means for subgroups with less than 10 members were not calculated.

Table 11
Means, Standard Deviations and Number of Examinees
by Level and Language Background
for the Pre-CPT and the CPT
(in Pre-CPT/CPT Scale Scores)

PRE-CPT MEANS TABLE

LEVEL	LISTENING			READING			STRUCTURE		
	Mandarin	OthrChin	English	Mandarin	OthrChin	English	Mandarin	OthrChin	English
HIGH SCHOOL									
	--	97.00	72.29	--	97.64	75.97	--	96.55	76.91
2nd Year	--	27.68	19.45	--	27.68	15.74	--	29.52	21.22
	--	(11)	(34)	--	(11)	(31)	--	(11)	(32)
	129.22	106.69	83.59	115.15	103.79	80.49	124.00	108.96	87.43
3rd Year	18.60	24.16	20.95	19.37	23.20	18.32	16.68	23.22	22.64
	(27)	(72)	(69)	(27)	(72)	(68)	(27)	(72)	(69)
	139.80	123.79	81.22	118.30	121.56	86.94	119.00	130.00	92.03
4th Year	20.21	17.69	28.65	39.69	17.84	21.49	25.75	20.25	26.03
	(10)	(34)	(36)	(10)	(34)	(36)	(10)	(34)	(36)
COLLEGE									
	133.90	107.19	88.25	105.32	96.05	90.37	114.33	104.40	92.93
2nd Sem	18.70	21.63	21.30	21.04	22.20	19.48	22.96	26.96	21.73
	(41)	(75)	(147)	(41)	(77)	(148)	(40)	(75)	(140)
4th Sem	--	--	103.46	--	--	108.83	--	--	112.58
	--	--	19.90	--	--	24.76	--	--	26.76
	--	--	(24)	--	--	(24)	--	--	(24)

CPT MEANS TABLE

LEVEL	LISTENING		READING		STRUCTURE	
	Chinese	English	Chinese	English	Chinese	English
	119.18	92.11	102.18	89.00	108.82	89.83
Beginning	20.00	15.47	18.78	14.08	27.34	12.77
	(17)	(389)	(17)	(363)	(17)	(391)
	111.21	98.12	105.36	98.56	111.97	97.48
Intermediate	18.52	15.04	19.29	15.35	21.73	15.34
	(39)	(934)	(39)	(932)	(39)	(935)
	133.00	111.10	134.00	115.11	146.27	111.12
Advanced	16.89	16.41	20.48	18.62	32.08	18.65
	(26)	(634)	(26)	(639)	(26)	(639)

To the degree that the means presented in Table 11 conform with expectations, Table 11 gives some indication of how well the equating was accomplished. For example, the means show that within any level, there is a wide divergence in performance on the CPT between Chinese and English speakers (as expected), and, for the Pre-CPT, that of Mandarin and non-Mandarin speakers of

Chinese. The means also reveal that for both English speakers and Chinese speakers, the two tests show consistent progress as level of instruction increases. As could be expected, mean scores on the Pre-CPT for second semester college students were very close to the mean scores on the CPT for the beginning level students across all subtests. Also, as might be expected, fourth (and especially third) year high school students do not do quite as well as second semester college students. On the whole, there are very few unexpected results in Table 11, and these may be due to the small number of examinees contributing to the means of some of the groups. It should also be kept in mind that default settings for BIGSTEPS produced scale score estimates for many of the native speaking Pre-CPT examinees who had perfect scores. This has no doubt inflated the means for the native-speaking Pre-CPT examinees.

Summary and Conclusions

The paper demonstrated the ability of the Rasch model computer program BIGSTEPS to vertically equate a scale of Chinese language proficiency using a concurrent calibration method. In comparisons with an anchoring method, concurrent calibration appeared to have certain advantages, in addition to its simplicity. In studies on all the items based on the test data, we found only very minor differences between the item difficulty logit values produced by the two methods and those produced when no equating was involved. However, in studies involving only the common items, we saw that the concurrent calibration method had a beneficial effect on the calibration of the common items. Further comparisons may highlight other differences in the outcomes when using the two methods. On the basis of this study, we would encourage users of BIGSTEPS to employ the concurrent calibration method in test equating.

ENDNOTES

1.One does not have to be able to read Chinese to respond to questions in the listening section.

2. It should be remembered that the CPT Structure items were the hardest on the test for the CPT population. Although the easiest items were selected for equating purposes, their lack of variance as a set and their lack of appropriateness for the Pre-CPT population, which is much lower in ability than the CPT population, undoubtedly was the reason for the much lower correlation between the logit values.

REFERENCES

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Linacre, J.M., & Wright, B.D. (1991a). BIGSTEPS: Rasch-model computer program. Chicago: MESA Press.

Linacre, J.M., & Wright, B.D. (1991b). A user's guide to BIGSTEPS. Chicago: MESA Press.

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