DOCUMENT RESUME

ED 377 699 FL 022 663

AUTHOR Coniam, David
TITLE Designing an Ability Scale for English Across the Range of Secondary School Forms.
PUB DATE Sep 94
NOTE 8p.; For complete volume, see FL 022 657.
PUB TYPE Reports - Descriptive (141) -- Journal Articles (080)
JOURNAL CIT Hong Kong Papers in Linguistics and Language Teaching; n17 p55-61 Sep 1994

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Networks; Databases; English (Second Language); Foreign Countries; Instructional Program Divisions; *Item Banks; Item Response Theory; Language Skills; *Language Tests; Rating Scales; Secondary Education; Second Language Instruction; *Test Construction; Test Items
IDENTIFIERS *Hong Kong

ABSTRACT

The development of a language ability scale for English as a Second Language in Hong Kong secondary schools is described. The project used a local computer network designed as a database of teacher-designed English language tests. From test items in the database, a series of seven tests was prepared, one for each grade level. Tests for the lowest three grades and upper four grades were prepared and pilot tested separately. Tests were piloted in mid-range-ability schools, in two classes at each grade level, providing a sample size of about 500 for each test. Based on the results, a common scale was developed and validated for the seven tests. Development of the item bank continues. (MSE)

Designing an ability scale for English across the range of secondary school forms.

David Coniam
Dept. of Curriculum Studies, The University of Hong Kong.

This paper details the construction of a common scale which attempts to span the English language ability range of students in the Hong Kong secondary school system. The TeleNex Computer network project, which operates out of the University of Hong Kong, aims to provide English language teachers in Hong Kong secondary schools with professional support. One of these areas of support is a testing database, which is attempting to recycle teacher-produced tests. To refine and feed tests back into the system at points appropriate to the ability levels of other schools and classes, the necessity for a common scale became apparent. Tests with common items were therefore designed and administered to Secondary forms 1 - 7. Using Item Response Theory techniques, the common items were then used as the basis for the common scale.

Introduction

This paper describes the initial construction of a common ability scale for English language across the Hong Kong secondary school system. The project operates under the aegis of the TeleNex Computer network out of the University of Hong Kong.[1] This network aims to provide English language teachers in Hong Kong secondary schools with professional help and input via computer to their teaching, as well as supporting them by means of a number of databases. One of these databases is a testing database, which was established with the objective of supplying teachers with a variety of reliable tests at various levels across the secondary school age and ability ranges.

The TeleNex testing database therefore aims to recycle analysed and refined teacher-produced tests at appropriate points throughout the database. To be able to insert teachers' tests at different points in the database in such a manner, however, it became apparent that a scale calibrated across the range of the Hong Kong secondary school system needed to be established as a reference point for different levels of ability. The author had worked with secondary school teachers, getting them to design tests, and had discovered that many teachers' concepts of test difficulty and the intended target audience were often very disparate. For example, a test which was destined for a 'low ability' Secondary 4 class of, for example, appeared to be rather more suitable - after item analysis had been conducted - for a 'high ability' Secondary 5 class. It was therefore decided to set up an item bank of short items which could then be drawn upon selectively. These could be appended to a test, and students' performance on these items would be used to determine the level of the remainder of the test they had taken.

[1] TeleNex operates from the Department of Curriculum Studies at the University of Hong Kong. It was set up with a donation of some $4 million in funding for both hardware and personnel from the Hong Kong Telecom Foundation.

The limitations of using classical item analysis as an instrument for comparing performance of groups or tests have received considerable attention in the literature on testing and measurement over the past decade. (Henning, 1984, 1987; Wright and Stone, 1979; Woods and Baker, 1985.) Classical methods of measurement theory are limited to one particular sample, essentially by using the correlation coefficient as the principal statistic. While classical methods are acceptable for single groups which can be normatively evaluated, or compared, it is not easy to translate these methods to situations where a number of tests need to linked together - as in the formation of an item bank, or where the ability of subjects needs to be calculated, independent of their scores on one particular test. See Henning (1984) for a comprehensive overview of the disadvantages of classical methods of measurement theory and the advantages of item response theory.

Since the bank of items needed to be established for students of widely differing samples of ability, and for samples of students who needed be compared one with another, Item Response Theory (IRT) appeared to be the appropriate measurement procedure to adopt in the current research project. The standard unit in IRT is the *logit*, which is a subject's log odds of producing a correct response to a particular test item.

Methodology

Seven tests - one for each form - were then designed with common items across adjacent forms. For forms Secondary 4-7, the tests were prepared with items from previous Hong Kong Examination Authority (HKEA) materials, and assigned to tests on the basis of their facility values in the public examinations. Discrete-point multiple-choice (m.c.) items were selected as the item type to be used - for a number of reasons. The HKEA was prepared to allow TELEC access to a substantial number of items (some 1,500 items were culled from various past HKEA papers); this meant that a bank of items could be established in a reasonably short period of time. Secondly, once the bank has been established, such short items would intrude only minimally into the time required to administer the main test. Typically, a short m.c. item takes a student 30 seconds to answer: 10 items would therefore only take away 5 minutes of the actual testing time. The marking - and subsequent analysis - of short m.c. items is also less time-consuming than other item types.

For forms Secondary 1-3, items had to be specifically prepared and pre-tested. Each test consisted of between 40 and 60 'proprietary' items for a given form, as well as items which were 'common' to the forms above and below. This elaborate design of common items across three levels was instituted in order that the validity of the scale could be examined from more than one angle. The test for Secondary 4, for example, consisted of a total of 80 mc items. 50 of these 80 items were proprietary S4 items; that is, the 50 S4 items were only found in the S4 test. The other 30 items comprised 10 common S4 items, 10 common S3 items and 10 common S5 items.

The tests were initially administered to mid-range ability schools in the pilot set-up of the TeleNex project, i.e. omitting schools which could be defined as extremely weak or extremely able.[2] Each school was asked to run the tests on two mid-ability level classes at each form. This gave a sample size of between around 500 subjects for each test.

[2] The definition of extremely able or extremely weak was made on the basis of a school's results in the Secondary 5 HKCEE public examination.

Initial analysis of test resul..

It is often assumed (this is the view of the HKEA) that the 'ideal' mean of the facility values of items on norm-referenced tests is generally around .50, since this ensures greater discrimination among the subjects taking the examination. Although facility values are not at issue here, these can be viewed as an approximate guide as to how a test fits its intended sample. Table 1 below presents the mean of the common items intended for each form.

Table 1: Test analyses: mean facility values of common items

SUBJECTS	S1	S2	S3	S4	S5	S6	S7
S7						.43	**.32**
S6					.58	**.36**	.26
S5				.57	**.43**	.29	.21
S4			.56	**.49**	.35		
S3	.59	.47	**.47**	.35			
S2	.58	**.45**	.44				
S1	**.51**	.33					

ITEMS

[Figures in larger font size cross-reference the common items for a specific form.]

It was mentioned above that test items were selected on the basis of appropriate fit for a particular form. It can be seen, however, that the fit is not as exact as might have been expected. This can in part be attributed to the fact that running the tests in November is quite early in the year, since the school year has only been underway for 2 or 3 months and for forms other than Secondary 5 and 7 there are still another 6 months of the school year left to run, with the public examinations typically beginning each year in April.

Taking .50 as the preferred test mean, the majority of the set of common items appear to be functioning reasonably well. The common item means for Secondary forms 1-4 are all close to a mean of .50, so given that the students will improve slightly in the remainder of the school year, the items have not been too badly targeted. The S6 and S7 test means have emerged rather lower than would have been hoped for at .36 and .32.

In terms of differentiating between different forms' performance the items appear to be functioning well. It can be seen that with all sets of items, the form below have found the items more difficult and the form above have found them easier. Consider the S4 set of items for instance. The S4 students achieved a mean of .49 on the 10 S4 common items. The S3 students found this set of items considerably more difficult with a mean of .35, while the S5 students found them much easier, with a mean of .57.

The only case where item means did not differentiate in the manner in which it had been intended was with the S2 and S3 tests. The means of both sets of common items were very similar for both groups of students. The S2 students achieved a mean of .45 on the S2 items and a mean of .44 on the S3 items. This matched quite closely with the results achieved by the S3 students who

had a mean of .47 on both tests. However, the fact that the S4 students found the S3 items easy and the S1 students found the S2 items difficult suggests that the potential discrepancy here may not lie as much with the items as with the students. This matter will be examined in more depth below.

The data was then analysed using a one-parameter IRT model. Logit values are computed with a mean of zero, and a standard deviation of 1. To avoid having negative values in the results logit values were rescaled to a normative mean of 60, with a spacing factor of 9.1. Wright and Stone (1979, p. 191ff) note that a scale without negative values is generally easier to interpret than one with negative values. This is especially the case with the current project where teachers are eventually to be referred to a suggested point on the scale according to a set of scores recorded by their students.

60 was selected as the additive constant since test items rescaled in this manner will generally have values of between 50 and 100. Since the scale is to be used with teachers, it was felt that the similarity between these figures and percentages would make the statistics of the system slightly less off-putting. The figure of 9.1 is used as a multiplicative constant following from Wright's discussion of the desirability of 'user-friendly' rescaled units (Wright, 1977, p. 203).[3]

The results of the seven tests are presented in table 2 below.

Table 2: Test analyses: mean logit values of common items

SUBJECTS	S1	S2	S3	S4	S5	S6	S7
S7						58.4	**63.0**
S6					51.0	**60.1**	64.8
S5				52.0	**57.5**	63.6	67.3
S4			56.7	**58.3**	65.6		
S3	56.5	58.5	**59.0**	63.9			
S2	64.9	**59.9**	60.9				
S1	**61.1**	69.0					

ITEMS

[Figures in larger font size cross-reference the common items for a specific form.]

The common items had been carefully selected on the basis of their match to the ability of a particular form. Given this, it had been hoped that the mean of each set of common items would have a mean close to 60.0. On the whole, the items appear to have been generally on target. The S7 items, however, with a mean of 63.0, suggest that the items are slightly above the average ability of the S7 sample, while the S5 items would appear to be slightly below the average S5 ability.

[3] Wright adopts the figure of 9.1 since this figure allows for the interpretation of the results of the probability of a person succeeding at a particular item in terms of user-friendly 'regular' intervals (.10, .25, .50 etc).

Going from left to right across the scores for a particular form, it will be noted - as with the mean facility indices of the common items from table 1 - that lower (i.e. easier) scores have been obtained by all forms. Consider again the scores of the S6 group of subjects. The S6 students' scores on the S6 common items centre on 60.1 suggesting that these items closely fit the ability of the target group. The S6 students' score of 51.0 on the S5 common items suggests quite correctly that their ability is above that of the S5 group. Conversely the S6 students' score of 64.8 on the S7 items suggests that S6 ability is substantially below that of the S7 group.

The discrepancy with the closeness of the S2 and S3 groups' scores in terms of facility values on the common items of both the S2 and the S3 tests can again be observed in the logit values of the common items of these two tests. Again, however, the S1 students have found the S2 items difficult while the S4 students found the S3 items difficult. The conclusion that is therefore drawn is that there is little difference between the abilities of these two forms.

Constructing the common scale

The next step was to construct a common scale. While the common items for the various levels appeared to be performing appropriately for their particular level, it was decided to use all the proprietary items for a given level. This larger number would reduce any possible skewing of the results, which might happen, for example, if a common item which was substantially easier or more difficult than its partners. At this stage, misfitting items were also removed. With the S1 test, then, the starting point of the scale, the item mean was computed with 64 rather than 10 items.

The S1 scores were taken as the starting point, since values would then rise. If S7 is taken as the starting point, the scale may well end in negative values. While this is technically unimportant in that the values are arbitrary, it was mentioned above that it is easier to interpret scales which do not have negative values rather than ones which do.

The scale was constructed vertically; that is, by anchoring one subject's set of scores on a test with the test above it. For example, the set of common items in the S1 test had a mean of 60.0. The S2 test was therefore analysed with the S1 common items in the S2 test set at 60.0. From this, the values for the S2 common items on the S2 test were obtained - 67.2. These values for the S2 common items were then set in the S3 test before the S3 test was analysed.

The scale is presented in table 3 below. (It will be recalled that the scale has a normative mean of 60.0 and a spacing factor of 9.1. These units are now referred to as TAAS values (TeleNex Average Ability Scores).)

Table 3: linked TAAS values of common items

S7	88.8
S6	84.9
S5	78.7
S4	70.7
S3	68.4
S2	67.2
S1	60.0

The scale that has emerged shows a range of 28.8 TAAS points - 3.16 logits. This is comparable to the range of ability described by Henning (1984) with reference to subjects enrolled on a pre-university intensive English programme at five different levels in the US. Henning's scale showed a mean person ability span of some 2.8 logits. The range of student ability in the current sample would appear to be greater than that in Henning's sample since the sample extends across seven forms as opposed to five levels. The scale might therefore be expected to come out as rather longer than that produced by Henning. This is in fact the case, although only by approximately half a logit (some 5 TAAS points).

The scale does not, however, show a totally linear progression across the forms. Between S1 and S2 there is a substantial difference in ability: almost one logit (9 TAAS points) of difference. Between S2 and S3, as has been remarked, there appears to be only a minimum amount of difference, with the S3 group appearing marginally more able than the S2 group.

Between S3 and S4, S4 and S5, S5 and S6, and S6 and S7, the ability difference then appears to be more regular, with approximately half a logit between each form.

A validation of the scale presented in table 3 can be observed if the scores are now compared horizontally with the values obtained from the common items in table 2. This compares different forms' scores on the same items rather than using, e.g. the Secondary 2 items to set the values for the Secondary 3 items. Consider the S5 items which have been taken by four forms. These items show a range of 15.3 TAAS points: 52.0 (by the S4 subjects) to 67.3 (by the S7 subjects). This compares reasonably accurately with the values derived from table 3, where the S4 to S7 scale shows a range of 16.3 TAAS points. The situation with the four sets of S3 items taken by the S1 to S4 forms is not quite as exact, with a range of 10.4 TAAS points on the horizontal item comparison as against 12.9 on the vertical scale. This differential represents approximately one quarter of a logit, which appear acceptable, given the range of four forms.

Conclusion and significance

The basic scale which has emerged, that is of students in roughly the mid-ability ranges of each form, extends some 3.33 logits. This compares well with the results reported by Henning (1984) where a slightly more restricted scale of learners of English gave a mean scale extending some 2.8 logits. The current sample would be expected to have a rather larger range due to the fact that the subjects are drawn from the across the entire Hong Kong secondary school system.

As a step towards calibrating tests submitted by teachers to the TeleNex testing database and to suggesting appropriate entry points for students of differing abilities, the scale in the current research project would appear to be a viable construct: Secondary forms 1 to 7 can all be placed on a common scale.

Nonetheless, the scale must be seen as a preliminary measure. While the basic scale has been devised from a sample of some 500 subjects per test, this is still too small to claim it is representative of the Hong Kong school population as a whole. As a first step, further exploration needs to be done by examining, for example, the weaker end of Secondary 1 as well as the more able end of Secondary 7. This will then give a clearer picture of the more extended range of ability.

The current research exercise contributed some 500 items to the item bank. If the item bank is to be in general circulation in Hong Kong, however it needs to be substantially expanded. Millman and Arter (1984, p. 319) suggest that a rule of thumb is roughly a factor of 10 to the number of items that could be used on any one occasion. While it is difficult to quantify exactly the number of users who may be accessing the database at different levels, it appears that an item

bank in the region of some 2,000 items may be appropriate. Currently, further items are being trialed with a view to expanding the item bank to a size approaching this dimension.

The Education Commission, following the recommendations in the recent report of the Hong Kong Education Department's *Working Group on Support Services for Schools with Band 5 Students* (1993) is now investigating English Language proficiency at the lower levels of attainment, i.e. Band 5 schools, in the Secondary school system. A calibrated scale with a sufficiently large item bank would be a extremely valuable resource in this regard.

Acknowledgements

The author would like to thank the Hong Kong Examination Authority for access to past examination material. He would also like to thank Michael Milanovic and Neil Jones of the University of Cambridge Local Examination Syndicate for advice and help with test design and analysis.

References:

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing* 1, 123-133.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Millman, J. and Arter, J.A. (1984). Issues in item banking. *Journal of Educational Measurement* 21, 315-330.

Woods, A. and Baker, R. (1985). Item Response Theory. *Language Testing* 2, 119-140.

Working Group on Support Services for Schools with Band 5 Students. (1993). *Final Report.* Hong Kong: Hong Kong Government.

Wright, B.D. (1977). Solving measurement problems with the Rasch Model. *Journal of Educational Measurement* 14, 97-116.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design*.

David Coniam has been involved in English language teaching and teacher education for 20 years. He has taught in the U.K., Spain, Iran, Saudi Arabia and Hong Kong. He is currently on leave from the Chinese University of Hong Kong, being on attachment as a lecturer to the Teachers of English Language Education Centre at the Department of Curriculum Studies, University of Hong Kong. His main interests are computer analysis of language, language testing and methodology.