言語テスティングにおけるラッシュモデルのアイテムバンキングへの応用
Application of Rasch Measurement to Item Banking in Language Testing

中村　優治 NAKAMURA, Yuji

● 慶應義塾大学
Keio University

Keywords 言語テスト，アイテムバンキング，ラッシュモデル，項目反応理論，多肢選択テスト
language testing, item bank, Rasch model, Item Response Theory, Multiple-choice test

ABSTRACT

ラッシュモデルの具体的応用例として考えられるのが，アイテムバンキングへの活用である．アイテムバンクは項目反応理論の分析手法を用いて統計処理を行った項目（アイテム）を蓄積したもので大規模な言語テストの開発のみならず，クラスルームにおける身近なテスト作成・テスト結果の考察にも威力を発揮するものである．本稿では理論的側面を概観したのち，50項目のテストデータを用いて，アイテムバンクの構築手順ならびにその長所，短所をのべ，さらに，実際の読解力テスト項目構築時の問題点などを指摘し今後の研究への示唆を行う．

The present research mainly deals with the basic idea of item banking: 1) how the items are calibrated for the storage, 2) how the persons' abilities are measured, 3) what the advantages and limitations of item banking are, and 4) problems of the Item Response Theory in language testing. Item characteristics can be determined either by traditional item statistics (called Classical Test Theory) or a newer method of estimating item statistics called Item Response Theory. This paper takes Item Response Theory for item characteristics. Furthermore, the sample data for the present research is taken from a multiple-choice test, which consists of 50 items and was conducted to 809 students.

1. Introduction

Since the 1960s there has been a growing interest in item response theory (IRT), a term which covers a range of models used to score tests. All of these models assume that a given test is unidimensional (Fulcher and Davidson, 2007).

Using the item response theory test designers can create a scale, usually scaled from around -4 to +4, upon which all items or tasks can be placed, and the value on the scale is the item difficulty. Test takers who take the items can be placed on to the same scale, and their scores are interpreted as person ability. As such, there is a direct connection between ability and difficulty (Fulcher and Davidson, 2007).

Henning (1989) says that latent trait measurement or item response theory refers primarily, but not entirely, to three families of analytical procedures. They are: the one-parameter (or Rasch Model), the two-parameter, and the three-parameter logistic models. The first parameter is a scale of person ability and item difficulty; the second parameter is a continuous estimate of discriminability; the third parameter is an index of guessing.

The Rasch model is an item response theory (IRT) one-parameter model developed by Geroge Rasch, which states that the probability of a correct response is a function of the difficulty of the item and the ability of the candidate. The term one-parameter refers to the item difficulty parameter (Davies et al, 1999). The model makes it possible to predict the likelihood of a correct answer to a given test item on the basis of the knowledge of two variables: item difficulty and person ability (Fulcher and Davidson, 2007).

The advantage of a Rasch analysis is that it can provide sample-free, scale-free measurement, that is to say scaling that is independent of the samples or the tests/questionnaires used in the analysis (Wright and Masters, 1982; Linacre, 1989). The Rasch model can also provide the researcher with information on how to organize the test items in terms of level of difficulty, spread of item difficulty, test length etc. in order to obtain optimal precision of measurement.

Among the applications of the Rasch model, item banking is a useful one for language testing. Item banking is creating a pool of items with known and invariant measurement characteristics. The Rasch model provides estimates of item difficulties which are meaningful irrespective of ability level tested. In a Rasch analysis, different tests can be formed into an overlapping chain through the employment of anchor items, which are common to adjacent forms. The forms can be targeted to particular groups of learners, yet linked into a common scale (Council of Europe, 2001).

This paper mainly focuses on how the model can contribute to the idea of item banking in terms of language testing. It also indicates the limitations of the Rasch model in the field of language testing.

2. Purpose of the research and research design

The present research mainly deals with the basic idea of item banking: 1) how the items are calibrated for storage, 2) how persons' abilities are measured, 3) what the advantages and limitations of item banking are, and 4) problems of the Item Response Theory in language testing.

Item characteristics can be determined either by traditional item statistics (called Classical Test Theory) or a newer method of estimating item statistics called Item Response Theory. This paper takes Item Response Theory for item characteristics. Furthermore, the sample data for the present research is taken from a multiple-choice test, which consists of 50 items and was conducted to 809 students.

2. 1 Subjects

809 freshman university students in K University

2. 2 Materials/ Instruments

A placement test for measuring students' English reading ability as well as grammar and vocabulary knowledge was used. It had four components: a grammar section (15 items), a vocabulary section (10 items), a reading section (3 long passages with five questions each: 15 items) and a cloze section (10 items) for a total of 50 items.

N.B. The reading section had three reading passages which were classified as beginners' level, intermediate level, and advanced level in terms of the content, the topic, and the vocabulary level out of the teachers' teaching experience. The length of the passages were about 400-500 words. Also the purpose of the cloze section was to measure students' grasping ability from the context.

3. Theoretical background and rationale

Beeston (2000) states that an item bank is a large collection of test items that have been classified and stored in a database so that at a later time, they can be chosen for new tests. The items are all classified according to certain characteristics such as the topic of a text, the testing point for an item as well as statistical information about item difficulty. It is important that all of the item difficulties have been located on a common scale of difficulty so that any combination of items can be put into a new test and the item difficulties added together to give a precise measure of the difficulty of that test.

Gronlund (1998) also indicates that item banks are files of various suitable test items that are coded by subject area, instructional level, instructional objective, and various pertinent item characteristics (e.g. item difficulty and discriminating power. Item banks are commonly used for the construction of equivalent or alternate forms of standardized tests (different combinations of homogeneous items are drawn from the bank), and as the basis for computer adaptive tests (items at a suitable level of difficulty for individual candidates are retrieved from the computer bank as required).

Choppin (1979) describes an item bank as a large collection of test questions organized and catalogued like the books in a library. The idea is that the test user can select test items as required to make up a particular test. Since one would think in terms of item banks with several thousand items, the number of possible tests which could be composed from such a bank is enormous. He claims that the great advantage of this system is its flexibility. Tests can be long or short, easy or difficult at will.

According to Davies et al (1999), the requirements for an item bank are 1) an adequate pool of test items, 2) an inventory of the abilities and content which each item purports to measure, 3) statistical data indicating the characteristics of each item as evidenced in test trialing (e.g. item difficulty and item discrimination indices), and 3) a theory or construct of ability which enables the meaning of scores on any test which may be constructed from the banked items to be interpreted. They further suggest that latent trait models are particularly useful in item banking because they have the advantage of allowing item scores to be translated into estimates of ability on a common scale. Thus, all tests deriving from a logit scale item bank are automatically equated since a person's score on any combination of test items can be converted into an ability estimate on the common bank scale. This means that any group of people can be given a test made up of items particularly suitable for them, yet all the results can be compared to one another.

Hozayin (2000) proposes three important characteristics of item banks: 1) storage, 2) coding and 3) item characteristics (difficulty and discrimination). Firstly, item banks are stored in files. Secondly, the coding and classifying of items is essential at both the storage stage and the retrieval stage. Items are classified according to the subject area, the instructional level etc.

Furthermore, Wright and Bell (1984) claim that the definition of an item bank is beyond storage and coding. An item bank is not just a collection of items but a bank of carefully calibrated test items. To calibrate items means to standardize them and make them more precise. In the process of increasing precision, we need to investigate item characteristics (item difficulty and item discrimination).

When items are calibrated and joined to a common bank of items, any cluster of these items can be used to measure ability that would be located on the same scale as ability measured by any other cluster of these items. This is called test-free person measurement. In other words, because items have been calibrated for difficulty it is possible to select items to match the known ability range of the examinees.

4. Practical procedures of item calibration and person measurement through Rasch calibration for item banking

Rasch calibration applies a probabilistic model to data in order to construct linear measures. These measures are not only linear but they are also accompanied by relevant estimates of their statistical validity and precision. This greatly enhances our information concerning the measure of the persons and the calibration of the items.

It is impossible to estimate a finite ability for persons who correctly answer all or none of a set of items. Where a person is labeled as a misfit this is not pejorative. IRT applies a probabilistic model to the actual data. If the model cannot account for the data, a person or item is flagged as misfitting. What this means is that an instance of person misfit can usually be attributed to anomalous test-taking behavior of some kind (Baker, 1997; Fulcher and Davidson, 2007). In such cases all we know is that these persons are more (or less) able than this test can measure. Thus, the first step in calibration involves setting aside persons with extreme scores (cf. Bode and Wright, 1999).

Let us take a look at our sample data in Table 1 below. In this table one student out of 809 should be set aside because of extremely good score (in this case this student all got 50 items correct). (See Appendix for the terms in the table. Also, cf.Linacre and Wright,1998 and 2001.)

Similarly, it is impossible to estimate a finite difficulty for items that are answered correctly by all (or none) of the persons taking them. Then all we know is that these items are too easy or too difficult for this sample of persons. Data editing also sets aside items with extreme scores. (cf. Bode and Wright, 1999).

Again let us look at our sample data in Table 2 below. In this table there are no items which should be set aside because of extreme scores.

Since cases with extreme scores have been removed (1 in persons, and none in items in the sample data), the data for the remaining persons and items are used in the following analysis.

In order to free these persons and item scores from sample size and test length, they are transformed into proportions of their maximum possible values. To linearize these proportions, they are converted to log odds, or logits (usually from –3 to 3), by taking the natural log of the proportion incorrect for items or failures for persons. This transforms the proportions to a linear scale (Bode and Wright, 1999).

Logit scores (person ability and item difficulty) are further transformed into measure scores on a 0-100 scale in the present research, which should be more familiar to the readers to understand the test data. Also we can avoid negative scores of low achievers and easy items.

Accordingly, the aforementioned two tables (Table 1 and Table 2) provide us with item-free person ability measures and person-free item difficulty measures. Table 1 shows item-free person ability measures, while Table 2 indicates person-free

Table 1 STUDENT STATISTICS: MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REA S.E.	INFIT		OUTFIT		PTMEA CORR.	Student
					MNSQ	ZSTD	MNSQ	ZSTD		
715	50	50	106.8	18.4	MAXIMUM ESTIMATED MEASURE					10508048
44	48	50	87.0	8.0	1.15	.4	3.37	1.6	-.05	10516690
16	47	50	82.4	6.6	1.12	.4	1.03	.4	.14	10503437
493	47	50	82.4	6.6	1.15	.5	1.37	.7	.09	10511474
172	46	50	79.0	5.5	.85	-.2	.45	-.6	.40	10512503
354	46	50	79.0	5.5	.79	-.4	.55	-.4	.41	10515745
527	46	50	79.0	5.5	.92	-.1	.66	-.2	.32	10500957
536	46	50	79.0	5.5	.88	-.2	.62	-.3	.36	10506233
623	46	50	79.0	5.5	.81	-.3	.40	-.7	.43	10501412
7	45	50	76.3	5.0	1.00	.1	.85	.0	.27	10516031
178	45	50	76.3	5.4	1.16	.6	.87	.1	.20	10514478
279	45	50	76.3	5.0	.92	-.1	.69	-.2	.35	10501585
481	45	50	76.3	5.0	.89	-.2	.62	-.4	.38	10507121
608	45	50	76.3	5.0	1.02	.2	1.09	.4	.23	10511998
790	45	50	76.3	5.0	.95	.0	.84	.0	.31	10503017
132	42	47	75.4	5.2	1.06	.3	1.48	.8	.18	10413659
33	44	50	74.0	4.6	.83	-.5	.58	-.6	.44	10515522
120	44	50	74.0	4.6	.88	-.3	.83	-.1	.37	10516633
136	44	50	74.0	4.6	.89	-.3	.97	.2	.35	10515366
204	44	50	74.0	4.6	.72	-.9	.41	-1.0	.53	10502536
350	44	50	74.0	4.6	.90	-.2	.80	-.1	.36	10501804
429	44	50	74.0	4.6	.94	-.1	.71	-.3	.36	10515247
585	44	50	74.0	4.8	1.07	.3	1.10	.4	.24	10505072
students in between are omitted for the convenience sake										
809	16	50	40.8	3.8	1.27	1.6	1.21	.9	.24	10501018
188	15	50	39.6	4.0	1.35	2.0	1.62	2.0	.13	10508126
580	15	50	39.6	3.4	.90	-.5	.83	-.6	.51	10511511
799	15	50	39.6	4.0	1.38	2.1	1.52	1.7	.13	10512873
399	14	50	38.4	4.4	1.58	2.9	1.85	2.4	-.05	10308606
512	14	50	38.4	3.5	.91	-.5	1.21	.8	.46	10516711
701	14	50	38.4	4.1	1.38	2.0	1.69	2.1	.09	10501980
187	10	40	37.4	4.5	1.29	1.4	2.70	3.4	.05	10516871
728	13	50	37.1	4.3	1.40	2.0	1.88	2.3	.07	10504704
253	12	50	35.8	4.3	1.37	1.8	1.58	1.6	.11	10513313
282	12	50	35.8	4.1	1.23	1.2	1.26	.8	.23	10514255
335	12	50	35.8	4.5	1.51	2.3	1.77	1.9	.00	10507827
407	12	50	35.8	3.7	.78	-1.2	.64	-1.1	.60	10517471
712	12	50	35.8	4.0	1.20	1.1	1.56	1.5	.21	10516072
577	11	50	34.4	3.9	1.09	.5	1.34	1.0	.30	10407833
MEAN	33.1	49.8	59.2	3.6	1.00	.0	.98	.0		
S.D.	6.6	1.2	7.8	.7	.14	.9	.30	1.0		

Table 2 ITEMS STATISTICS: MEASURE ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REAL S.E.	INFIT		OUTFIT		PTMEA CORR.	Item
					MNSQ	ZSTD	MNSQ	ZSTD		
47	126	800	77.8	1.1	1.10	1.5	1.44	4.1	.09	47CL
49	209	789	70.7	.9	1.03	.7	1.19	3.1	.23	49CL
20	305	807	64.8	.8	.97	-1.0	.97	-.8	.35	20V
35	308	806	64.7	.8	.96	-1.4	.98	-.5	.36	35Rb
30	326	808	63.6	.8	.97	-1.0	.99	-.2	.35	30Ra
11	331	806	63.3	.8	1.14	5.5	1.22	6.1	.14	11G
44	357	797	61.6	.8	1.05	2.3	1.10	3.2	.25	44CL
16	386	806	60.2	.8	1.10	4.3	1.14	4.5	.20	16V
17	388	806	60.1	.8	1.03	1.5	1.06	2.1	.28	17V
42	406	803	59.0	.8	1.02	.9	1.04	1.5	.30	42CL
43	427	806	57.9	.8	1.00	.2	1.01	.4	.32	43CL
22	433	802	57.5	.8	1.01	.7	1.01	.4	.31	22V
19	437	806	57.3	.8	1.09	4.0	1.15	4.9	.20	19V
25	443	808	57.1	.8	1.04	1.8	1.05	1.5	.28	25V
4	446	807	56.9	.8	1.07	2.8	1.09	2.8	.24	4G
9	458	806	56.2	.8	.98	-1.0	.96	-1.3	.36	9G
32	476	808	55.2	.8	.98	-.8	.97	-.8	.35	32Rb
21	479	807	55.0	.8	.95	-1.8	.93	-2.0	.39	21V
41	483	804	54.7	.8	1.04	1.5	1.03	.8	.28	41CL
18	498	804	53.8	.8	1.03	.9	1.07	1.9	.29	18V
5	509	804	53.2	.8	1.02	.7	1.00	.1	.31	5G
33	511	808	53.2	.8	.99	-.3	.99	-.2	.34	33Rb
23	520	807	52.6	.8	1.03	1.0	1.02	.5	.29	23V
31	530	806	51.9	.8	1.03	1.0	1.05	1.2	.28	31Rb
38	530	802	51.8	.8	1.01	.3	1.00	.1	.31	38Rc
48	533	797	51.4	.8	1.00	-.1	1.03	.6	.32	48CL
50	538	788	50.7	.8	1.06	1.8	1.07	1.5	.25	50CL
12	556	806	50.3	.8	1.00	-.1	1.06	1.3	.31	12G
15	568	806	49.5	.8	1.00	.1	1.00	.1	.31	15G
36	567	799	49.2	.8	1.03	.7	1.02	.5	.28	36Rc
6	578	807	48.9	.8	.93	-1.8	.89	-2.1	.40	6G
37	579	805	48.7	.8	.95	-1.3	.90	-1.9	.38	37Rc
3	597	806	47.5	.8	.87	-3.2	.81	-3.6	.47	3G
24	628	808	45.3	1.0	1.14	2.9	1.20	2.9	.12	24V
8	628	805	45.1	.9	1.03	.6	1.06	.9	.26	8G
45	634	803	44.5	.9	.94	-1.3	.96	-.6	.37	45CL
1	658	808	42.7	.9	.93	-1.2	.90	-1.2	.36	1G
40	669	804	41.4	1.0	.96	-.7	.93	-.7	.33	40Rc
7	683	807	40.2	1.0	.89	-1.8	.79	-2.3	.41	7G
46	682	801	39.8	1.0	.91	-1.3	.79	-2.3	.39	46CL
39	700	803	38.0	1.1	.93	-.9	.79	-2.0	.36	39Rc
26	705	808	37.9	1.1	.98	-.2	.95	-.5	.28	26Ra
2	708	808	37.5	1.1	1.00	.0	.96	-.3	.26	2G
34	711	807	37.1	1.1	.90	-1.2	.77	-2.1	.38	34Rb
13	710	806	37.0	1.1	.87	-1.7	.92	-.7	.40	13G
27	735	808	33.8	1.3	.98	-.1	.85	-1.1	.27	27Ra
29	735	808	33.8	1.3	.88	-1.3	.60	-3.4	.41	29Ra
14	748	807	31.3	1.4	.91	-.8	.73	-1.9	.34	14G
10	757	807	29.4	1.5	.93	-.6	.71	-1.8	.31	10G
28	788	807	19.1	2.3	.97	-.1	.67	-1.2	.20	28Ra
MEAN	534.3	804.6	50.0	.9	.99	.2	.98	.2		
S.D.	149.8	4.3	11.2	.3	.06	1.7	.15	2.1		

item difficulty measures. For example, in Table 1, Student 132 has a Rasch ability measure 75.4 (although there are other students with the same measure in this present data), which is an estimate of this person's ability regardless of which items they responded to. This means that it is not necessary for every test taker to take every item in a pool in order to ensure that the item statistics are meaningful. Another example is Student 577 who has a Rasch ability measure 34.4 and this is the lowest ability among these 808 measured students.

In Table 2, for example, Item 47 has an item calibration or difficulty measure of 77.8 which is an estimate of this item's difficulty regardless of the ability level of the persons who responded to it. Another example is Item 28 which has an item calibration or difficulty of 19.1 and this is the easiest item among these 50 items.

In addition, Rasch analysis provides two estimates of misfit: infit and outfit. Infit is sensitive to irregular patterns of responses for items close to a person's ability level. Outfit is sensitive to unexpected responses to items far from the person's ability level. Both are useful indicators of potential problems. Large outfit indicates the presence in the data of unexpected off-target responses. Large infit, in contrast, indicates a central pattern of response incoherence. Although overfit or small misfit values provide insight into how an item set might be shortened by deleting redundant items, they are generally not a concern (Bode and Wright, 1999). Therefore, we can be entirely flexible about misfit.

Let us examine Table 3 for the present sample test data. This table shows the 10 calibrated items in the misfit order.

A rule of thumb for the acceptable range of infit and outfit scores in multiple choice questions is between 0.7 and 1.3. If this is a high stakes test, which is used to make a very important or critical decision about someone's future, for example, we strictly stick to this rule. Since 1.44 in Item 47, 0.67 in Item 28, and 0.60 in Item 29 are beyond the range, these three items should, accordingly be taken out from the item list. However, in the present research which is to demonstrate the item banking procedure, considering the number of the items is small and the percentage does not seem fatal to the analysis (3 out of 50 items or 6 % of the whole test), we can leave them as they are in this list. Thus, Rasch measurement not only estimates item difficulties and the precision of these estimates but also tests the fit of each item to the construct implied by the set of items. Then, in addition to estimating person measures, it examines the response patterns of persons to determine whether they are responding as expected.

After items are calibrated according to item response theory (IRT) or the Rasch one parameter model in the present research, they can be stored in an item bank according to a common metric of difficulty. This is generally true regardless of the equality of the ability or size of the subsequent person samples tested although there is an expected number of test takers to make a generalization. The item bank becomes more than just a catalog of used items with descriptions of their successes and failures. It becomes an ever-expanding test which spans the latent ability continuum beyond the measurement needs of any one individual, but which may be accessed to gather items appropriate to any group of persons from the same general population with respect to the ability measured (cf. Henning, 1987).

All items or tasks that survive review and examination need to be banked, or stored, in a format that allows easy retrieval according to any number of search criteria that may be used as identifiers. These criteria are usually those used in describing tasks for test specifications. However, the item bank should also contain any statistical data that are associated with an item or task, such as its facility value, discrimination index. This allows test assemblers

Table 3 ITEMS STATISTICS: MISFIT ORDER

ENTRY NUMBER	RAW SCORE	COUNT	MEASURE	REAL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEA CORR.	Item
47	126	800	77.8	1.1	1.10	1.5	1.44	4.1	A .09	47CL
11	331	806	63.3	.8	1.14	5.5	1.22	6.1	B .14	11G
24	628	808	45.3	1.0	1.14	2.9	1.20	2.9	C .12	24V
49	209	789	70.7	.9	1.03	.7	1.19	3.1	D .23	49CL
19	437	806	57.3	.8	1.09	4.0	1.15	4.9	E .20	19V
16	386	806	60.2	.8	1.10	4.3	1.14	4.5	F .20	16V
44	357	797	61.6	.8	1.05	2.3	1.10	3.2	G .25	44CL
4	446	807	56.9	.8	1.07	2.8	1.09	2.8	H .24	4G
18	498	804	53.8	.8	1.03	.9	1.07	1.9	I .29	18V
50	538	788	50.7	.8	1.06	1.8	1.07	1.5	J .25	50CL
17	388	806	60.1	.8	1.03	1.5	1.06	2.1	K .28	17V
12	556	806	50.3	.8	1.00	-.1	1.06	1.3	L .31	12G
8	628	805	45.1	.9	1.03	.6	1.06	.9	M .26	8G
31	530	806	51.9	.8	1.03	1.0	1.05	1.2	N .28	31Rb
25	443	808	57.1	.8	1.04	1.8	1.05	1.5	O .28	25V
42	406	803	59.0	.8	1.02	.9	1.04	1.5	P .30	42CL
41	483	804	54.7	.8	1.04	1.5	1.03	.8	Q .28	41CL
23	520	807	52.6	.8	1.03	1.0	1.02	.5	R .29	23V
36	567	799	49.2	.8	1.03	.7	1.02	.5	S .28	36Rc
48	533	797	51.4	.8	1.00	-.1	1.03	.6	T .32	48CL
5	509	804	53.2	.8	1.02	.7	1.00	.1	U .31	5G
22	433	802	57.5	.8	1.01	.7	1.01	.4	V .31	22V
43	427	806	57.9	.8	1.00	.2	1.01	.4	W .32	43CL
38	530	802	51.8	.8	1.01	.3	1.00	.1	X .31	38Rc
15	568	806	49.5	.8	1.00	.1	1.00	.1	Y .31	15G
2	708	808	37.5	1.1	1.00	.0	.96	-.3	y .26	2G
30	326	808	63.6	.8	.97	-1.0	.99	-.2	x .35	30Ra
33	511	808	53.2	.8	.99	-.3	.99	-.2	w .34	33Rb
27	735	808	33.8	1.3	.98	-.1	.85	-1.1	v .27	27Ra
26	705	808	37.9	1.1	.98	-.2	.95	-.5	u .28	26Ra
32	476	808	55.2	.8	.98	-.8	.97	-.8	t .35	32Rb
35	308	806	64.7	.8	.96	-1.4	.98	-.5	s .36	35Rb
9	458	806	56.2	.8	.98	-1.0	.96	-1.3	r .36	9G
20	305	807	64.8	.8	.97	-1.0	.97	-.8	q .35	20V
28	788	807	19.1	2.3	.97	-.1	.67	-1.2	p .20	28Ra
40	669	804	41.4	1.0	.96	-.7	.93	-.7	o .33	40Rc
45	634	803	44.5	.9	.94	-1.3	.96	-.6	n .37	45CL
21	479	807	55.0	.8	.95	-1.8	.93	-2.0	m .39	21V
37	579	805	48.7	.8	.95	-1.3	.90	-1.9	l .38	37Rc
1	658	808	42.7	.9	.93	-1.2	.90	-1.2	k .36	1G
6	578	807	48.9	.8	.93	-1.8	.89	-2.1	j .40	6G
39	700	803	38.0	1.1	.93	-.9	.79	-2.0	i .36	39Rc
10	757	807	29.4	1.5	.93	-.6	.71	-1.8	h .31	10G
13	710	806	37.0	1.1	.87	-1.7	.92	-.7	g .40	13G
14	748	807	31.3	1.4	.91	-.8	.73	-1.9	f .34	14G
46	682	801	39.8	1.0	.91	-1.3	.79	-2.3	e .39	46CL
34	711	807	37.1	1.1	.90	-1.2	.77	-2.1	d .38	34Rb
7	683	807	40.2	1.0	.89	-1.8	.79	-2.3	c .41	7G
29	735	808	33.8	1.3	.88	-1.3	.60	-3.4	b .41	29Ra
3	597	806	47.5	.8	.87	-3.2	.81	-3.6	a .47	3G
MEAN	534.3	804.6	50.0	.9	.99	.2	.98	.2		
S.D.	149.8	4.3	11.2	.3	.06	1.7	.15	2.1		

to find items that meet specific criteria (Fulcher and Davidson, 2007).

Item banking is usually done electronically so that searching through the bank is done easily. This means that it is necessary to maintain the system and update the database with new items and tasks as they are produced and approved for operational use (Fulcher and Davidson, 2007).

5. Advantages of item banks

Hozayin (2000) says that a main advantage of calibrated item banks is in the ease of test development. A set of items, in the form of a test, may be withdrawn from the bank, and the teachers will know how difficult this set of items is for the test takers. The teachers will also know how well these items discriminated between the students who have learned the target content and those who haven't. Additionally, Hozayin (2000) claims that a second advantage of calibrated item banks is that they can provide the basis for a curriculum map, in which the learning objectives included in the curriculum are ordered by difficulty. This will allow teachers to gain greater insight into the learning process of their students, to confirm that what they think is difficult or easy is actually difficult or easy for the students. Therefore, it will be much easier to chart the progress of individual students over time (cf. Choppin, 1979).

Wright and Bell (1984) describes an advantage of item banks from the viewpoint of students in the following way. A well constructed item bank can provide the basis for designing the best possible test for every purpose. This is because it is not necessary for every student to take the same test in order to be able to compare results. Students can take the selections of bank items most appropriate to their levels of development. The number of items, their level and range of difficulty, and their type and content can be determined for each student individually, without losing the comparability provided by standardized tests. Comparability is maintained because any test formed from bank items, on which a student manifests a valid pattern of performance, is automatically equated, through the calibration of its items onto the bank, to every other test that has been or might be so formed.

Furthermore, Wright and Bell (1984) also point out an advantage of item banks from the viewpoint of teachers. A well organized item bank enables teachers to construct a wide variety of tests. They need not settle for standard grade level tests or administer the same test to every student in a class, school. They can consider who is to be measured and for what purpose and select items accordingly. They can tailor each test to their immediate educational objectives without losing contact with the common core of bank items. They can write, bank and use new items that reflect their own educational goals while retaining, when their new items fit the bank, the opportunity to make whatever general comparisons they may require.

It is also important to note that because all of the items drawn from a particular bank are calibrated onto one common scale, teaches can compare their test results with one another, even when their tests contain no common items (Wright and Bell, 1984). This opportunity to compare results quantitatively enables teachers to examine how the same topic is learned by different students working with different teachers and hence to evaluate alternative teaching strategies. With common curriculum strands as the frames of reference, it becomes possible to recognize subtle differences in the way school subjects are mastered. The investigation of which teaching methods are most effective in which circumstances can become an ongoing, routine part of the educational process. In other words, tests constructed from item banks can promote an exchange of ideas, not only about assessment, but also about curricula (Wright and Bell, 1984).

6. Limitations of item banks

As with any approach to educational measurement, there are limitations on item banks. Using an item bank will not eliminate the need for the test developers to evaluate the quality of the items stored in the bank. In addition, the test developers must be sure that the content tested by the item reflects the target content (Hozayin, 2000).

Furthermore, Choppin (1979) says that it is important to realize that item banking is not the final solution to all the problems posed by educational assessment. No item bank can be better than the material that is put into it, and users of assessment materials will continue to carry responsibility for ensuring that their tests are fair, appropriate, reliable and valid. An item bank should be a living thing with test materials being added and the classification system updated as new developments occur either in our understanding of the subject matter or in teaching practices (Choppin, 1979).

7. Some problems of the Item Response Theory and Rasch Measurement for L2 reading assessment in Language Testing

Theoretical papers emphasize the multidimensionality of the reading construct, whereas descriptions of testing practice speak to the need for unidimensional scores, particularly for placement (Chapelle and Douglas, 2006).

One issue is about a rather new technological test method called the Computer Based Testing (CBT) or the Internet Based Testing (iBT). With this testing format a variety of measures of reading ability can be quickly administered, such as reading rate, word recognition, vocabulary knowledge, reading fluency (Grabe, 2000). Grabe (2000), in addition, says that this test also offers a variety of texts along with integrated reading tasks across multiple texts.

Reading passage and sub-questions are mostly dependent, while Grammar and Vocabulary questions in discrete-point tests can be independent. Items in Gap filling tests are all inter-dependent. The issue is the unidimensionality and local independence in each sub test of a whole English proficiency test.

Even in an independent reading text, each passage has a couple of sub-questions, and they are usually very interdependent. Let alone, in an integrated test where not only main idea questions but also word recognition or cohesive or coherence questions are asked like in an iBT TOEFL, it is almost impossible to claim the local independence of each item.

8. Conclusions

Item response theory facilitates item banking by allowing all of the items to be calibrated and positioned on the same latent continuum by means of a common metric. Also, it permits additional items to be added subsequently without the need to locate and retest the original sample of examinees. Furthermore, an item bank permits the construction of tests of known reliability and validity based on appropriate selection of item subsets from the bank without further need for trial in the field (Henning, 1987)

Hozayin (2000) stresses the point that a carefully developed item bank may serve as the basis for adaptive testing which is usually called computer adaptive testing (CAT), (since adaptive tests are almost always delivered on a computer). This allows item selection to match the specific ability level of the individual student who is taking the test. Accordingly, as Wright and Bell (1984) suggests, using this adaptive capacity of item banks, teachers need not settle for standard grade level tests or administer the same test to every student in a class or school. They can consider who is to be measured and for what purpose and select items accordingly. They

can tailor each test to their immediate educational objectives without losing contact with the common core of bank items.

Finally, we have learned how the item calibrations and the person measurement are conducted using the Rasch model for item banking. The idea of item banking along with the improvement in computer technology will lead to a new way of language test development and use, even though there may be some hurdles to be cleared in our future process.

References

Baker, R. (1997). *Classical Test Theory and Item Response Theory in Test Analysis*. Language Testing Update Special Report No. 2. Lancaster: University of Lancaster.

Beeston, S. (2000). *UCLES Research Notes 2*. University of Cambridge Local Examinations Syndicate.

Bode, R. K. and Wright, B. D. (1999). *Rasch measurement in higher education. Higher Education: Handbook of Theory and Research*, vol.XIV. (pp.287-316).

Chapelle, A. C. and Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge UK: Cambridge University Press.

Choppin, B. (1979). Testing the questions--the Rasch model and item banking. In M. St. J. Raggett, C. Tutt, P. Raggett (Eds.). *Assessment and Testing of Reading: Problems and Practice*. London: Ward Lock Educational.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching*, assessment. Cambridge: Cambridge University Press.

Davies, A, A. Brown, C. Elder, K. Hill, T. Lumley and T. McNamara. (1999). *Studies in Language Testing 7: Dictionary of language testing*. Cambridge UK: Cambridge University Press.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: an advanced resource book*. London: Routledge.

Grabe, W. (2000). *Reading research and its implications for reading assessment*. In A. Kunnan (Ed.), Fairness and validation in language assessment (pp.226-62). Cambridge: Cambridge University Press.

Gronlund, N. (1998). *Assessment of Student Achievement*. 6th Edition. Needham Heights, MA: Allyn and Bacon.

Henning, G. (1987). *A Guide to Language Testing*. New York, Newbury House Publishers.

Hozayin, R. (2000). Item Banks: Definition and Development. Paper presented at the Sixth EFL Skills Conference at the American University in Cairo, January 25th–27th, 2000.

Linacre, J. M. (1989). *Multi-faceted Measurement*. Chicago, IL: MESA Press.

Linacre, J. M. and Wright, B. D. (1998, 2001). *A User's Guide to Bigsteps/ Winsteps: Rasch-Model Computer Program*. Chicago, IL: MESA Press.

Wright, B. D. and Bell, S. R. (1984). Item banks. What, why, how. *Journal of Educational Measurement*. 21(4),331-345 Winter.

Wright, B. D.and Masters, G. N.(1982). Rating Scale Analysis. Chicago, IL: MESA Press.

Appendix

INFIT is an information-weighted fit statistic, which is more sensitive to unexpected behavior affecting responses to items near the person's ability level.

1) MNSQ is the mean-square infit statistic with expectation 1.
2) ZSTD is the infit mean-square fit statistic standardized to approximate a theoretical mean 0 and variance 1 distribution.

OUTFIT is an outlier-sensitive fit statistic, which is more sensitive to unexpected behavior by persons on items far from the person's ability level.

1) MNSQ is the mean-square outfit statistic with expectation 1.
2) ZSTD is the outfit mean-square fit statistic standardized to approximate a theoretical mean 0 and variance 1 distribution.

REALSE

1) SE is the standard error computed over the persons or over the items.
2) REALSE is computed on the basis that misfit in the data is due to departures in the data from model specifications (cf. Linacre and Wright, 1998).