International Conference on Education and Educational Psychology (ICEEPSY 2012)

Examining Quality of Mathematics Test Items Using Rasch Model: Preliminary Analysis

Nordin bin Abd. Razak*, Ahmad Zamri bin Khairani, Lei Mee Thien

*School of Educational Studies, 11800 Universiti Sains Malaysia

Abstract

Producing a high quality of test items needs a substantial time and efforts in developing and examining each test item. Literature suggested this limitation can be remedied with a large collection of test items along with their measurement characteristics which termed as item bank. However, a well-developed item bank is formed by a set of high quality of test items using a rigorous development and validation procedures. In achieving this, this study attempts to describe the development of Malaysian secondary school Form two Mathematics item bank; and to examine the psychometric properties of the test items which developed based on Malaysian Mathematics Curriculum Specifications. The sample consisted of 1 316 Form two students selected from 12 secondary schools in the state of Penang and Perak. Data was analyzed using Rasch Unidimensional Measurement Model with WinSteps computer program. The results showed that 19 out of 160 items were excluded as not fitting Rasch measurement model. Thus, only 141 items were stored in the item bank and empirically fulfilled the quality as suggested by the Rasch measurement model. This study contributes to create Malaysian Mathematics item bank which deserves to be further created, tested and modified the computer adaptive testing (CAT) program.

Keywords: Mathematics item bank, item development, item validation, Rasch Measurement Model

1. Introduction

Test is a powerful educational tool that should give an accurate picture of students’ knowledge and skills in the subject area or domain being tested. As such, accurate measure of students’ achievement is very crucial for curriculum and instruction planning and educational-related program evaluation. Conversely, test that underestimate students’ actual knowledge and skills cannot serve these important purposes. With this regards, there is a dire need to develop good quality of test items in terms of item development and validation.
However, to accomplish this task is not easy as producing a good quality test items demand a lot of time, effort, and energy. In addition to the capability in enhancing subject matter knowledge, teachers need to spend relentless hours for tests planning including the process of assembling items, writing, and determining the difficulty level of each item (Richichi, 1996; Ahmad Zamri, 2010). This perhaps serve as one the possible reasons to explain why most test items developed by teacher are considered fail to discriminate between high and low ability; and items are not function according to their intended used (Richichi, 1996).

In addressing this pitfall, literature suggests that much of the burden of test construction can be reduced when a large collection of good quality of test items is available to either teachers or test developers (e.g., Millman & Arter, 1984). A large collection of test items along with their measurement characteristics is termed as item bank or interchangeably recognized as item pools. Item bank is a notion that is widely defined, from a loose and unrestricted definition such as ‘any collection of test questions’ (Millman, 1984, p.85) to ‘a composition of coordinated questions that develop, define, and quantify a common theme and thus provide an operational definition of a variable’ (Wright, 1984, p.211). Meanwhile, in a more detail manner, Choppin (1981), Rudner (1998) refer item bank as a large collection of good test items for which their quality is analyzed and known, and which are systematically stored in a computer so that they are accessible to students and teachers for measuring their achievement or ability.

Intriguing, Rudner (1998) highlights the application of item bank, namely, item banking provides substantial savings of time and energy in developing test due to its convenience in depositing and storing items. There are other two main reasons to explain the sufficient application of item bank. First, items in an item bank can be edited, withdrawn, and deposited (Rudner, 1998). Secondly, its capability to develop several test with the characteristics of flexibility, security, and consistency. The reasons mentioned convince item banking to be a useful tool for educational system to monitor educational achievement (MacCann & Stanley, 2009). Considering this, a well-developed item bank is only indicative through good quality of test items as listed follows according to Educational Testing Service (1992).

(i) Items must be fair, valid and reliable in order to create fair, valid and reliable tests;
(ii) A test is only as good as each item on it. If items don’t really measure the standard, the test results will not be useful;
(iii) Instructional decisions should be made on the basis of valid assessments;
(iv) Teachers need good data to make better use of their limited instructional time; and
(v) Older and repurposed items often do not address today’s instructional standards.

As such, an item bank involves a thorough process of items filtering and calibrating before item storing (Wright & Bell, 1984) in developing good quality of items. Statistically, calibrating items means items need to be standardized for the purpose of precision. In simplicity, item bank should be developed and validated with accurate assessment in item level in order to produce a set of good quality items.

Remarkably, a robust calibrated item bank that utilises Rasch Model provides numerous advantages to test developers such as flexibility, consistency, economy and security (Umar, 1999). Advancement in using sophisticated computer software further enhances the possibilities of development of a calibrated item bank. Computer program such as WinSteps for test development and FastTest for item banking facilitates the development item bank where evaluation of each item and the formation of each test can be much sufficient made.

A variety area of studies have successfully developed and validated their test using Rasch Model. For instance, Kazeem (1988) uses Rasch Model to construct achievement test while Abdel Fateh El Korashy (1995) utilizes Rasch Model to select items for mental test. Meanwhile, Barman and Boone (1997), Ludlow (2001) and Salzberger (2002) uses Rasch Model to validate education tests while Heinemann et al. (1997) validate the measurement of stroke scale in medical research. The other significant empirical studies includes Baylor, Yorkston, Miller, and Amtmann (2009), Forkman, Boecker, Wirtz, Eberle, Westhoffen, Schauerte, Mischke, and Norra (2009), Gothwal, Wright, Lamourex, and Pesudov (2009), Heinz, Smith, and Kassel, and (2009), as well as Muis, Winne, and Edwards (2009).
In a sharp contrast, in Malaysia, test development remains as one of the most tedious aspects in research (Siti Aloyah & Siti Rahayah, 2001). In fact, there remains less validity evidence of some Malaysian educational instruments (T Subahan, 2003). One of the possible reasons is due to the local rapid development in methodological and computerization advancement in comparison to western countries such as United States of America, United Kingdom, and Australia. This limitation motivates this study to initiative the development and validation of the Mathematics test items that form the item bank using rigorous statistical technique with advance computer software, namely, WinSteps. In achieving this, the purpose of this study is to describe the development of Malaysian secondary school Form two Mathematics item bank which developed based on Malaysian secondary school Form two Mathematics curriculum; followed by examining the psychometric properties of the test items.

2. Method

2.1. Participants

The targeted population is Malaysian secondary Form two students in the states of Penang and Perak. A total of 12 secondary schools with six from each state were selected using Probability Proportionate to Size procedure. Specifically, a total of 120 students were selected from each selected school using Systematic Random Sampling procedures. Overall, samples consisted of 1440 Malaysian secondary Form two students selected from 12 selected secondary schools.

2.2. Instrumentation & Data Analysis Procedures

This study employed quantitative approach which comprised the following distinct fundamental stages.

Stage 1  Definition of Content
The purpose of this stage is to define the content of an initial pool items. The content is in line with the learning outcomes specified by Malaysian Mathematics Curriculum Specifications and 10 specific objectives of Malaysian secondary Mathematics curriculum as shown in Table 1.

Stage 2  Instrument Development and Validation

The instrument was in the form of test let which consisted of 40 questions with four options of A, B, C, and D. In sum, ten test lets was developed coverage of 13 topics mentioned following standardized format of Test Blueprints circulated by Malaysian Ministry of Education (MOE). These 10 test lets were self-developed following scale development and validation procedures suggested by Thien, Nordin, and Hazri (2011) as illustrated in Figure 1.
Table 1

Objectives of Malaysian Mathematics Curriculum for Secondary Schools & Learning Areas

| Objective                                                                 | Main Area  | Learning Area                                                                 |
|---------------------------------------------------------------------------|------------|-------------------------------------------------------------------------------|
| (1) Understanding definition, concepts, laws, principles and theorems related to Numbers, Shapes and Space, and Relationships. | Number     | Topic 1: Directed Number  
                       |              | Topic 2: Squares, Square Roots, Cubes and Cube Roots                          |
| (2) Widen applications of basic fundamental skills such as addition, subtraction, multiplication and division related to Numbers, Shape and Space, and Relationships. | Shape and Space | Topic 6: Pythagoras Theorem  
                       |              | Topic 7: Geometrical Construction  
                       |              | Topic 9: Loci in Two Dimensions  
                       |              | Topic 10: Circles  
                       |              | Topic 11: Transformation  
                       |              | Topic 12: Geometric Solid |
| (3) Acquire basic mathematical skills such as:                            | Relationship | Topic 3: Algebraic Expression  
                       |              | Topic 4: Linear Equations  
                       |              | Topic 5: Ratios and Proportions  
                       |              | Topic 8: Coordinates  
                       |              | Topic 13: Statistics |
| • Making estimation and rounding;                                        |              |                                                                              |
| • Measuring and constructing;                                            |              |                                                                              |
| • Collecting and handling data;                                          |              |                                                                              |
| • Representing and interpreting data;                                    |              |                                                                              |
| • Recognising and representing relationship mathematically;              |              |                                                                              |
| • Using algorithm and relationship;                                      |              |                                                                              |
| • Solving problem;                                                       |              |                                                                              |
| • Making decision.                                                       |              |                                                                              |
| (4) Communicate mathematically;                                          |              |                                                                              |
| (5) Apply knowledge and the skills of Mathematics in solving problems and making decisions; |              |                                                                              |
| (6) Relate Mathematics with other areas of knowledge;                    |              |                                                                              |
| (7) Use suitable technologies in concept building, acquiring skills, solving problems and exploring the field of Mathematics; |              |                                                                              |
| (8) Cultivate mathematical knowledge and skills effectively and responsibly; |              |                                                                              |
| (9) Inculcate positive attitudes towards Mathematics; and                 |              |                                                                              |
| (10) Appreciate the importance and the beauty of Mathematics             |              |                                                                              |
Snowballing is one of the approaches that able to generate a large pool of items and refine the items at the same time based on subject matter knowledge of item developers (Kanji & Asher, 1996). In this study, eight Malaysian secondary Form two Mathematics teachers were invited to be the item developers. The item developers have teaching experience for the subject of Mathematics ranged from five to 17 years. The items were developed with bilingual version: English language and Malaysian national language using backward translation from English to Malaysian national language due to the concern of multicultural validity as pointed out by Kirkhart (1995). Multicultural validity refers to an instrument measures what it supposes to measure as understood by the respondents of a particular culture (Kirkhart, 1995).

Following this, item developers first generated as many items as possible based on the topic and learning areas provided. Item developers were then paired together to share their subject matter knowledge to analyze the suitability of the generated items. At this stage, a total of 480 items were generated and subjected to be sorted following the level of difficulty using Q Sort method. Q Sort was employed to rank a pool of items following the desired criteria that can be the best for measuring the learning areas of interest (Colton & Covert, 2007). In this study, a total of 480 items were sorted following level of difficulty with the criteria of Low, Intermediate, and High. The criterion of Low consisted of the categories of Knowledge and Comprehension; Intermediate comprised Application and Analysis; as well as High comprised Synthesis and Evaluation. The item developers were responsible to sort the initial pool of 480 items into the desired criterion respectively. Table 2 reveals that 480 multiple choice items are included for pilot testing whereas the remaining 80 items are found to be totally unsuitable. In sum, the level of item difficulty which represented by the criteria of Low, Intermediate, and High were in the ratio of 5 : 3 : 2. Next, the selected items were subjected to content validity.
(b) Content Validity & Item Refinement

The purpose of conducting item content validation is to make sure the self-developed items are able to capture the measure of a specific learning area appropriately (Anastasi, 1988). Four Malaysian Mathematics subject matter experts were invited to determine the extent that the items tap into the learning areas of interest. Item wordings were refined based on the comments and suggestions from the subject matter experts and students. No single item was excluded at the end of this stage.

(c) Test Design

A preliminary analysis was conducted on ten students sample to ensure the appropriateness of question content, wording, sequence, format, layout, and instruction. However, only four out of 10 bilingual version test were administered in a preliminary study.

Table 2

| Topic | Question   | Level of Difficulty | Low | Intermediate | High |
|-------|------------|---------------------|-----|--------------|------|
|       |            |                     | Knowledge | Comprehension | Application | Analysis | Synthesis | Evaluation |
| 1     | Q1-Q4      | 1                   | 19      | 2            | 12      | 6       |
| 2     | Q5-Q6      | 4                   | 6       | 4            | 9       |
| 3     | Q7-Q10     | 1                   | 13      | 10           | 8       | 6       |
| 4     | Q11-Q14    | 3                   | 13      | 15           | 1       | 5       |
| 5     | Q15-Q18    | 4                   | 21      | 12           | 2       | 2       | 2       |
| 6     | Q19-Q20    | 14                  | 2       | 4            | 4       |
| 7     | Q21-Q22    | 17                  | 2       | 1            | 1       |
| 8     | Q23-Q26    | 18                  | 4       | 11           | 5       | 1       |
| 9     | Q27-Q28    | 10                  | 1       | 1            | 7       | 6       |
| 10    | Q29-Q32    | 5                   | 5       | 5            | 12      | 10      |
| 11    | Q33-Q36    | 22                  | 2       | 5            | 6       | 3       |
| 12    | Q37-Q38    | 1                   | 9       | 1            | 1       | 2       | 4       |
| 13    | Q39-Q40    |                     | 20      |              |         |
| Sub-total |        | 15                  | 185     | 60           | 60      | 56      | 24      |
| Total |            | 200                 | 120     | 60           | 80      |
| %     |            | 50                  | 30      | 20           |

(d) Test Administration Procedure

The self-developed questionnaire was distributed personally to the selected secondary Form Two students with average age of 14 years. The questionnaire administration was supported with the informed consent from the Malaysian educational regulatory authorities. Participation was entirely voluntary and anonymous. The responsible teacher was required to read out the instructions to the students. Students have about 45 minutes to respond to the test let. The response rate is 91 percent with 1 316 sets of answer sheets were collected.

(e) Data Cleaning

Data cleaning refers to a process in ensuring the appropriateness codes for the optional answers (Thien et al., 2011). Data cleaning has to be conducted in order to avoid the presence of any out-of-range values which potentially create the extreme values of the data. In this study, data cleaning is used to determine whether data entry only restricted to the legitimate numerical or keycodes.
(f) Test and Item Calibration Design

An analysis need be carried out to ensure that the data fit the assumption of Rasch model in order to examine the measurement quality of the items. The psychometric properties of these test items were examined based on several indicators such as the point-item fit statistics indices, item and person reliability. The data were analyzed using WinSteps, a Rasch Unidimensional Measurement Model computer program so that all the item difficulties were ordered on the same linear scale along with the student measures of mathematical ability.

The Rasch model provides two fit statistics namely: infit and outfit Mean Square Statistics (MNSQ). The infit MNSQ is sensitive to unexpected responses to items near the person ability level and the outfit MNSQ is outlier sensitive. Mean square fit statistics are defined such that the model-specified uniform value of randomness is 1.0 (Wright & Stone, 1979). Person fit indicates the extent to which the person’s performance is consistent with the way the items are used by the other respondents. Item fit indicates the extent to which the use of a particular item is consistent with the way the sample respondents have responded to the other items. For this type of analysis, values between .70 and 1.30 logits (log odd units) are considered acceptable (Bond & Fox, 2001).

The psychometric properties of the test were also tested in terms of reliability and validity of the measures’ meaning and interpretation. Rasch analysis provides reliability indices for both item and examinee’s measure. High reliability for both indices are desirable since they indicate a good replication if the comparable items/examinees are employed.

3. Findings

Table 3 presents the findings based on four out of 16 tests. Initially, all tests consisted of 40 multiple choice items and the calibration was performed more than once with exception to Test 1 and Test 2. With regards to Test 1 and Test 2, the INFIT MNSQ values ranges from .80 and 1.23 and the OUTFIT MNSQ ranges from .74 and 1.30. These values lied within the acceptable ranges of fit statistics from .70 to 1.30. This indicates that all items in the tests working together to define student ability in Mathematics test. Meanwhile, for Test 3, three items are excluded from analysis to achieve the best item fit and 37 items were remained. With reference to Test 4, 16 items are excluded and only 24 items remained.

Table 3

| Test | Item Calibration based on Rasch measurement analysis |
|------|------------------------------------------------------|
|      | Initial Number of Item Measure Ranges Item Difficulty Person Infit MNSQ Outfit MNSQ Final No. of Item Person Item Reliability |
| 1    | 40 | -2.37 to 2.53 | 0.00 to 0.72 | 0.85 | 1.23 | 0.74 | 1.3 | 40 | 0.75 | 0.99 |
| 2    | 40 | -4.34 to 4.15 | 0.00 to -0.23 | 0.8 | 1.16 | 0.74 | 1.28 | 40 | 0.75 | 0.99 |
| 3    | 40 | -2.09 to 1.67 | 0.00 to -0.36 | 0.81 | 1.23 | 0.76 | 1.3 | 37 | 0.71 | 0.98 |
| 4    | 40 | -3.36 to 2.14 | 0.00 to 0.80 | 0.88 | 1.13 | 0.83 | 1.24 | 24 | 0.79 | 0.99 |

For Test 1, the item difficulties range from -2.37 logits \((SE=0.13)\) to + 2.53 logits \((SE=0.40)\) and for Test 2 the item difficulties range from -4.34 logits \((SE=.20)\) to +4.15 logits \((SE =.40)\). Item difficulties for Test 3 range from -2.09
logits ($SE=14$) and $+1.67$ logits ($SE=44$) and for Test 4 from $-3.36$ logits ($SE=22$) and $+2.14$ ($SE=37$). Test 1 ($M=72$, $SD=84$) and Test 4 ($M=80$, $SD=133$) seem to have higher students’ mean compared to item mean which is fix to 0.00. On the other hand, Test 2 ($M=-0.23$, $SD=82$) and Test 3 ($M=-0.36$, $SD=74$) were found to have lower mean from the item mean. This indicates the overall tests were slightly difficult in comparison to overall students’ ability. Meanwhile, the reliability of item difficulty measures in all tests was considered high ranged from .98 to .99. This suggests that the ordering of item difficulty was replicable with other comparable sample of examinees. Meanwhile, the parameter estimates of consistency of examinees’ measures, namely, Cronbach’s alpha was also considered high ranged from .71 to .79. The results showed the possibilities that the ordering of examinees proficiency can be replicated since most of the variance was attributed to true variance of the Mathematics proficiency construct. In addition, based on high Item Reliability indices, the result shows that there is a high consistency in ordering of the item if administered to comparable sample of students. Based on the abovementioned fit analysis, the descriptive statistics as well as the reliability indices, all items that have been kept demonstrate evidence of quality that useful in measurement of mathematical ability construct. They are ready to be use as tailored by the test users.

4. Conclusion

Based on the results, 19 out of 160 items were excluded as not fitting a Rasch measurement model and only 141 items were stored in the item bank and empirically fulfilled the quality as suggested by the Rasch measurement model. These items are coded according to competency area and instructional objective, as well as empirically derived measures of item difficulty obtained from the Rasch analysis. As such, it is particularly important to specify how items are expected to be ordered in difficulty. This provides a conceptual frame of reference for judging the validity and utility of the empirical ordering that subsequently derived from student performance. A well constructed item bank enables teachers to design the best possible test for every purpose. Teachers can tailor each test to their immediate educational objectives and consider who is to be measured without losing contact with the common core of bank items. 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. This study is an important step in creating Malaysian Mathematics item bank and further will be used to create, tested and modified the computer adaptive testing (CAT) program. However, additional analysis need to be carried out with other test booklets and test linking using common items in order to fit this purpose. In overall, it is important to realize that a well-planned and well-documented item bank is necessary for ensuring that the tests are fair, appropriate, reliable and valid and Rasch measurement may contribute to this process.

5. Acknowledgement

This article is made possible by the funding obtained from the Universiti Sains Malaysia

References

Abbdel Fateh EL Korashy (1995). Applying Rasch Model to the selection of items for mental ability test. Educational and psychological measurement, 5, 753-763.

Ahmad Zamri Kairani (2010). Application of the Bookmark Method in setting performance standards in Mathematics for Form 2 Students. Unpublished PhD thesis. International Islamic University Malaysia.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.
Baylor, C. R., Yorkston, K. M., Eadie, T. L., Miller, R., & Amtmann, D. (2008). The Levels of Speech Usage: A self-report scale for describing how people use speech. *Journal of Medical Speech-Language Pathology, 16*(4), 191-198.

Boone, W. J. (1997). Science attitudes of selected middle school students in China: A preliminary investigation of similarities and differences as a function of gender. *School Science and Mathematics, 97*(2), 96-103.

Choppin, B. (1981). Educational measurement and the item bank model. In C. Lacey & D. Lawton (Eds.), *Issues in evaluation and accountability*. London.

Colton, D., & Covert, R. W. (2007). *Designing and constructing instruments for social research and evaluation*. San Francisco: John Wiley & Sons, Inc.

Educational Testing Service (1992). *The Origins of Educational Testing Service*. Princeton, NJ: ETS.

Forkmann, T., Boecker, M., Wqirtz, M., Eberle, N., Westhofen, M., Schauerte, P., Mischke, K., & Norra, C. (2009). Development and validation of the Rasch-based depression screening (DESC) using Rasch analysis and structural equation modeling, *Journal of Behaviour Therapy and Experimental Psychiatry 40*(3), 468-478.

Gothwal, V. K., Wright, T. A., Lamoureux, E. L., & Pesudovs, K. (2009). Activities of daily vision scale? What do the subscale measures? *Investigating Ophthalmology & Visual Science, 51*(2), 694-700.

Heineman, A. W., Harvey, R. L., McGuire, J. R, Inberman, D., Lovell, L., Semik, & Roth, E. J. (1997). *Measurement properties of the NIH Stroke Scale during acute rehabilitation*. Retrieved Feb 18, 2007. http://stroke.ahajournals.org/cgi/content/28/6/1174.

Heinz, A.J., Smith, E.V., & Kassel, J.D. (2009). Caffeine Expectancy: Instrument Development in the Rasch Measurement Framework. *Psychology of Addictive Behaviors, 23*, 500-511.

Kanji, G. K., & Asher, M. (1996). *100 methods for total quality management*. Thousand Oaks, CA: SAGE.

Kazeem, A. M. (1988). *Application of Rasch model in the construction of achievement test in psychology*. Kuwait City, Kuwait. University Publication.

Kirkhart, K. E. (1995). Seeking multicultural validity: A postcard from the road. *Evaluation Practice, 16*(1), 1-9.

Ludlow, L. H., & Mahalik, J. R. (2001). Congruence between a Theoretical Continuum of Masculinity and the Rasch Model: Examining the Conformity to Masculine Norms Inventory. *Journal of Applied Measurement, 2*(3), 205-226.

MacCann, R.G. & Stanley, G. (2009) Item banking with embedded standards, *Practical Assessment Research & Evaluation, 14*(17), available online at http://pareonline.net.

Millman, J. and Arter, J. A. (1984). Issues in Item Banking. *Journal of Educational Measurement, 1*(4), 315-330.

Ministry of Education Malaysia. (2004). *Integrated Curriculum for Secondary School syllabus (Mathematics)*. Curriculum Development Center.

Muis, K. R., Winnie, P. H., & Edwards, O. V. (2009). Modern psychometrics for assessing achievement goal orientation:A Rasch analysis. British Journal of Educational Psychology, v79n3pp. 547-576.

Richichi, R. (1996, October). *An Item Response Theory Analysis of Multiple-Choice Items Chosen at Random From a Publisher's Test Bank*. Presentation at the 27th annual Northeastern Educational Research Association Conference, Ellenville, NY (ERIC Document Reproduction Service No. ED405367).
Rudner, L. (1998a). *Item Banking* (No. EDO-TM-98-04). USA: Eric Clearinghouse on Assessment and Evaluation, Washington, DC.

Salzberger, T. (2002). The illusion of measurement: Rasch versus 2-PL. *Rasch Measurement Transactions*, page 882.

Siti Aloyah Alias & Siti Rahayah Ariffin (2001). *Facing the challenges of school-based performance assessment. An evaluation of Peka Biologi programme*. Paper presented at the International Conference of Measurement and Evaluation in Education, Universiti Sains Malaysia, Penang, November.

T. Subahan Mohd Meerah (2003). *Development and validation of teacher evaluation instrument in Malaysia*. Paper presented at the International Conference of Measurement and Evaluation in Education, Universiti Sains Malaysia, Penang.

Thien, L. M., Nordin, A. R., & Hazri, J. (2011). Proceedings from International Conference on Measurement and Evaluation in Education. *Alternative procedures in scale development and validation*. ISBN-978-967-399-073-3.

Umar, J. (1999). Item Banking. In G. N. Masters, & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment*. New York: Pergamon Press.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review*, 3(1), 281-288.

Wright, B. D., & Bell, S. R. (1984). Item Banks: What, Why, How. *Journal of Educational Measurement*, 21, 331-345.

Wright, B.D. & Stone, M.H. (1979). *Best Test Design*. Chicago, ILL: MESA Press.