Bull. Nat'l Taiwan Nor. Univ.
Vol. 36. pp. 191-212 (1991)

師大學報
第36期 第191-212頁(民國80年)

APPLYING ITEM RESPONSE THEORY TO SCIENCE EDUCATIONAL MEASUREMENT:

GALT AND TIPS (II)

Rong-Fu Hsu

Department of Physics, NTNU

Abstract

This study developed the techniques for representing latent traits in Science Process Skills and Reasoning Ability measurement based on item response theory.

The three basic assumptions and specific objectivity in Rasch Model were examined in view of science education. Both linear logistic trait models and component latent trait models for accounting the nature of the abilities measured in test situations were also proposed.

Results of the study revealed that the predicted-observed patterns were affected by latent traits underlying each item and the stems for each item. The analysis might demonstrate the high sensitivity and resolution in detecting bias of test items of Rasch model in this study.

INTRODUCTION

The properties of the Rasch model and some item-response models are presented in this paper. The Rasch model and item-response models are accepted as powerful models for item analysis and theoretical basis for test-score interpretations. Based on the advance of cognitive psychology, some models for accounting for the nature of the abilities measured in test situations are proposed: for example, the linear logistic trait models (Fischer, 1983), component latent trait models (Whitely and Schneider, 1980), and many others.

In science education, there are only a few applications of these new test theories into educational and research designs. We tend to construct our tests according to classical test the-

ory, analyze the test items by traditional item-analysis techniques, and then complain that the theory does not satisfy our purposes. However, there are alternative choices after the estimation programs have been developed. Tests such as GALT, TIPS (II) and others may be analyzed by using the item-response theory.

RASCH MODEL

1. Basic Assumptions

The item-response models are based on the following assumptions:

(1) All items measure the same unidimensional trait: that is, there is only one ability underlying examinee performance;

(2) The function relating the probability of a correct response for an item to the underlying latent trait has a specified logistic form;

(3) Local stochastic independence holds: that is, the correct response to an item depends on the person's ability and the difficulty of the item, but not on which of the other items the person has previously answered (Bloomquist, 1984; van den Wollenberg, 1988; Hambleton and Swanminathan, 1985; Lord, 1980).

Under these assumptions, the item-response models can obtain sample-independent and item-independent measures, especially the Rasch Model will reach "specific objectivity" (Rasch, 1967; Andersen, 1977; Fischer, 1987).

2. The Simple Rasch model

Let X_{ag} be a random variable, $X_{ag}=1$ when person a answers item g correctly, and $X_{ag}=0$ for other cases. Let θ_a and δ_g denote the ability of person a and the difficulty of item g, respectively. According to the second assumption, the probability of person a answering item g correctly has the following form:

$$(1)\ P(X_{ag}=x_{ag} \mid \theta_a, \delta_g) = \lambda^{x_{ag}} / (1+\lambda),$$

where $\lambda = \exp(\theta_a - \delta_g)$. The person and item parameter are compensative. The person parameter is called nuisance parameter, and the item parameter is called structural parameter, since the former will generally influence the estimation of the later. Fortunately, this can be avoided by conditinoal maximum likelihood estimation approach.

Now, if there are N persons answering K items test, then according to local independency,

for person a,

$$(2)\quad P(X_{a1}=x_{a1} \ldots X_{ak}=x_{ak} \mid \theta_a, \delta_1 \ldots \delta_k) = \prod_{g=1}^{K} I\ P(X_{ag}=1 \mid \theta_a, \delta_g)$$

$$= \exp\left(\sum_{g=1}^{K} (\theta_a - \delta_g)^{x_{ag}}\right) \Big/ \left\{\prod_{g=1}^{K} (1+\exp(\theta_a - \delta_g))\right\} .$$

For all samples, the likelihood becomes

$$(3)\quad L = \prod_{a=1}^{N} P(X_{a1} \ldots X_{ak} \mid \theta_a, \delta_1 \ldots \delta_k)$$

$$= \exp\left(\sum_{a=1}^{N}\sum_{g=1}^{K} (\theta_a - \delta_g)^{x_{ag}}\right) \Big/ \left\{\prod_{a=1}^{N}\prod_{g=1}^{K} (1+\exp(\theta_a - \delta_g))\right\} .$$

Let $I_g \sum_{a=1}^{N} x_{ag}$, be the score of item g; $T_a = \sum_{g=1}^{K} x_{ag}$, be the person's raw score. The expression (3) becomes

$$(4)\quad L=\exp\left(\sum_{a=1}^{N} T_a * \theta_a - \sum_{g=1}^{K} I_g * \delta_g\right) \Big/ \left\{\prod_{a=1}^{N}\prod_{g=1}^{K} (1+\exp(\theta_a - \delta_g))\right\} .$$

Now, if we use (4) to estimate the values of θ_a and δ_g by maximizing likelihood L, this method is called unconditional or joint maximum likelihood estimation (Wright and Panachi-nathan, 1969; Hembleton, 1985; Fischer, 1980). The main advantages of the UCON or JML approach is that the estimation process can be easily implemented by the Newton-Raphsom method and the number of items to be analyzed can be up to hundreds. By maximizing (4),

$$(5)\quad (\partial / \partial \delta_g) \ln(L) = I_g - \sum_a (\exp(\theta_a - \delta_g) / (1+\exp(\theta_a - \delta_g))) = 0 \text{ and}$$

$$(6)\quad (\partial / \partial \theta_a) \ln(L) = T_a - \sum^{a} (\exp(\theta_a - \delta_g) / (1+\exp(\theta_a - \delta_g))) = 0$$

One theoretical problem of this method is that the estimation may be biased. As the number of items less than 10-15, the problem becomes more important (more detailed discussion of this problem, may be found in van den Wollenberg, et al., 1988 and Wright, 1988). Generally, the estimates of item parameters by UCON can be corrected by multiplying by factor (k-1)／k, where k is the length of test.

The conditional maximum likelihood approach is theoretically valid for RM (Andersen, 1970, 1972, 1977; Fischer, 1980; Gustafsson, 1980; van den Wollenberg et al., 1988). This method can eliminate the problem of UCON, but needs much more computational effort. The advantages of CML include theoretical completeness and statistcs for model fit the test. These

are introduced briefly in the following paragraph:

Let $\varepsilon = \exp(\theta)$ and $\xi = \exp(-\delta)$. For testee a who earns a raw score r,

$$(7)\quad P(r \mid \varepsilon_a, \xi_1 \ldots \xi_k) = \sum_r \prod_{g=1}^{K} (\varepsilon_a \xi_g)^{x_{ag}} / (1+\varepsilon_a \xi_g)$$

there are C (k,r) combinations for getting r raw score in k-items test. Define γ as

$$(8)\quad \gamma_r(E) = \sum_{T_a=r} \prod_{g=1}^{K} (\xi_g)^{x_{ag}},$$

where $E = (\xi_1 \ldots \xi_k)$, is the vector of item difficulty. γ_r is called the r-th elementary symmetry function (Fischer, 1983). By the help of γ, it is easy to obtain

$$(9)\quad P(r \mid \varepsilon_a, E) = (\varepsilon_a)^r \gamma_r / \prod_g (1+\varepsilon_a \xi_g).$$

(9) is the probability of getting raw score r given ability ε_v and item difficulties E. Divide (2) by (9).

$$(10).\ P(A \mid r,E) = \prod_g (\xi_g)^{x_{ag}}, / \gamma_r$$

where A is the response pattern ($X_{a1}=x_{a1}, \ldots, X_{ag}=x_{ag}$). For N testees, the likelihood becomes

$$(11)\quad L = \prod_{a=1}^{N} \{ \prod_{g=1}^{K} (\xi_g)^{x_{ag}}, / \gamma_r \}$$

It is important that the ability parameters are removed from (10) so that the estimation of item parameters can be sample free. The estimates of item parameters can be obtained by maximizing the ln (L) in (11).

$$(12)\quad T_g = \sum_{r=1}^{k-1} n_r \varepsilon_g \gamma^g_{r-1} / \gamma_r$$

where n_r is the number of testees who get r-items correct, and γ^g_{r-1} is the first derivation of γ_r with respect to ε_g.

The difficulty of CML is how to implement the computations of estimation processes. Fortunately, two algorithms were developed (Fischer, 1974; Gustafsson, 1980), i.e., the difference and summation of algorithms. By these procedures of estimation, the number of items analyzed by CML can be up to 100.

In summary, the advantages of the CML estimation approach include

1. that the nuisance (personal) parameters can be removed in the estimation procedure. This is important for obtaining sample free measurement.
2. that because of the theoretical completeness, there are sound statistics for model-fit tests

and flexibity for applications.

3. The modifications of the Rasch model

From the standpoint of modern psychology, the ability to solve test item is a very complex. For successfully solving a physics problem, for example, the testees are not only involved in trying to understand the verbal or non-verbal symbols, but also must have the correct physics concepts and some appropriate mathematics abilities. In other words, understanding what happens in test situations is really complex. This is why there are debates about the unidimensionality of the Rasch-family models. Bloomquist (1984) points out that the breadth of the latent trait underlying a test is a spectrum. In test construction, we always conceptualize that there is an objective for each item, and if all the items of a given test have common objective, then the test is unidimensional. This concept of unidimensionality seems plausible, but it rarely occurs in test situations. The assumption of strictly parallel forms in classical test theory implies the same concept of relations between objective and test items. Figure 1 presents two models of relations between test items and the assessment objective. Actually, the relations may be more complex than these. In the following paragraph, more complex test models derived from simple Rasch models are presented.

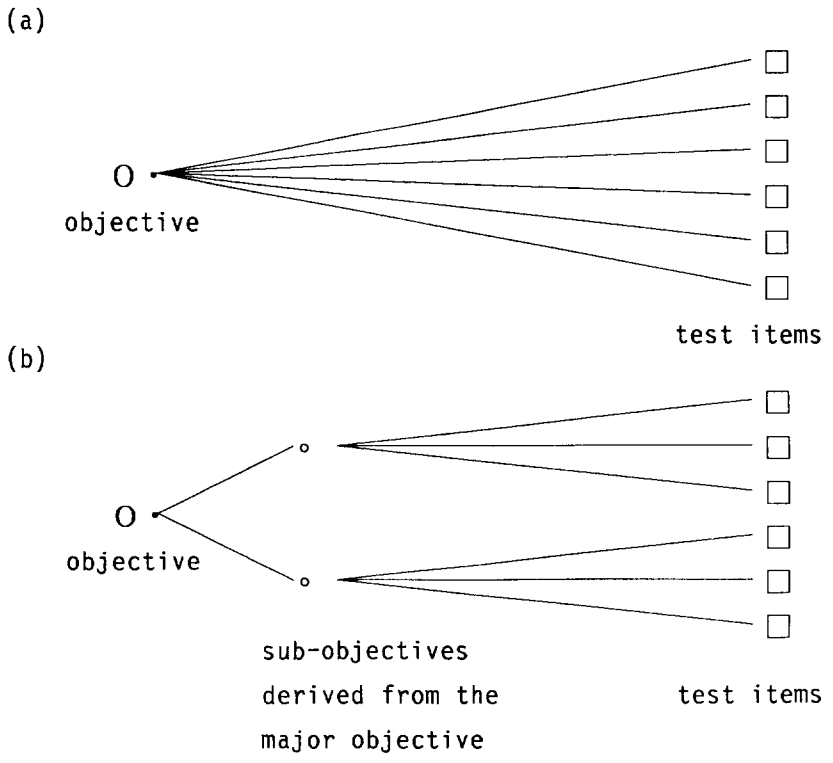


Fig 1. Two models for illustrating the relationship between test items and objective (s).

The LLTM--Linear Logistic Trait Models (Fischer, 1976, 1983, 1987; Spada and McGaw, 1985) are a general form of a simple Rasch model, which provides a method for decomposing item difficulty into components. This model is useful for understanding the contents of test items, and then helping test developers and users to establish their tests or interpret test results appropriately.

In LLTM. (1) can be rewritten as

$$(13)\quad P(X = 1 \mid \theta, \delta) = \exp(\theta - \delta) / (1+\exp(\theta - \delta))$$

with

$$(14)\quad \delta = \sum_{i=1}^{m} \eta_i * q_i + c.$$

Where c is a constant. The LLTM decomposites the item difficulty into m components, η_i is the parameter for the i-th component, and q_i is the hypothetical frequencies with which each component j influnences the solution of test items.

The estimation of parameters in LLTM can be implemented by CML approach (Fischer, 1983) and structural equation (Green and Smith, 1987), and the uniqueness of the solution in LLTM is shown by Fischer (1983). However, the estimation procedures in LLTM is more complex than in the simple Rasch model.

With some modification, the LLTM model can be flexibly applied to different test situations: for example, accounting for Piaget cognitive task performance (Spada and McGaw, 1985), cognitive complexity (Scheiblechner, 1972), intelligence test (Fischer and Forman, 1982), learning during testing, etc. The linear logistic models with relaxed assumptions (LLRA) are applicable for experimental design (Fischer, 1977).

The last test model introduced here is CLTM (component latent trait models, Whitely, 1980, 1984, 1985). The ration of CLTM is based on information processing theory. That is, ability is nothing but an inductive summary of weighted combinations of the underlying cognitive variables (Embretson, 1985). Two major cognitive variables are the level of knowledge and the number of options. The CLTM is process-product models; that is, it is assumed that solving the item requires the correct information outcome from several processing components. Therefore, this model is particularly applicable to complex intelligence test items such as verbal analogies, syllogisms, practical-judgement items, and mathematical-reasoning items.

The relationship of the component process products to the total item has been postulated by Whitely and Schneider (1980). which can be expressed as follows:

(15) $P(X_{ijT}=1) = (a\text{-}g) \prod_{k} P(X_{ijk}=1) + g.$

That is, the probability that the total item is solved $P(X_{ijT}=1)$ is the product of the subtask probabilities $P(X_{ijk}=1)$. Here a and g represent executive processing and guessing, respectively. The constant a is the probability that the component information is applied to the total task, given that correct outcomes have been obtained. The constant g is the probability that the correct outcome is obtained, given that at least one component is incorrect. The probability $P(X_{ijk}=1)$ can be the formula (1) mentioned above for Rasch model.

There are many other psychometrical models with potential value when applied to science educational measurenent, for example, the rule space model (Tatsuoka, 1987) and partial credit model (Master, 1982) may be useful to help us understand the nature of science abilities, and attitudes. But to what extent these models are available needs more empirical and theoretical study.

ANALYSIS OF GALT AND TIPS(II) BY SRM

This section presents two examples of using the Rasch model to analyze two tests used frequently in science educational research: the first is GALT (test of Group Assessment of Logical Thinking, Roaddrangka, Yeany and Padilla, 1983), which is designed to measure logical thinking ability by transfering the Piaget Tasks into multiple-choice tests. The second test is TIPS (II) (Test of Integrated Processes Skills, Burns et al., 1983), which is one of the most popular instruments for assessing integrated process skills. These two tests are traslated into Chinese with some modifications of item stems for ascertaining the equivalence between the original and translated tests. There are 21 and 34 items for the Chinese version of GALT and TISP (II), and about 2,000 grade 7 through grade 9 pupils were sampled for testing. The α coefficient of these two tests are .81 and .81, respectively. By factor analysis of the test data, the construct validity was examined (Lin, 1986; Lieu, 1988). Table 1 and table 2 list the contents the these two tests.

Table 1. The Objectives of GALT.

	Objective	Number of items
A.	Conservation	4
B.	Proportional reasoning	6
C.	Control varible	4
D.	Probabilistic reasoning	2
E.	Correlational reasoning	2
F.	Combinational reasoning	3

TABLE 2. The Classification of Items by Objectives of the Chinese Version of TIPS (II)

	Objective	Number of items
A.	Operational definition To identify a suitable operational definition of the described variable.	6
B.	Hypothesizing Given a description of investigaion or a problem, to identify a suitable hypothesis or one being tested	9
C.	Time-space graph Given a description of an investigation and obtained data, to identify a graph that represents the data, or given a graph of data from an investigation, and to identify the relationships between the variables	7
D.	Design an experiment	

	Given a hypothesis, to select a suitable design for an investigation to test it.	3
E.	Identify variable Given a description of an investigation, identify the independent, dependent, and controlled variables.	9

1. Analysis of GALT

In this paper, the results of analysis of GALT presented here are mainly to show the power of Rasch model, so only the first 10 items will be presented. these items included two objectives, 4 items for conservation and 6 items for proportional reasoning. For understanding the results presented below, we have to describe the item specifications in more detail. The first 4 items are weight, volume, length conservation and the independence of weight and volume (balls with the same volume raise water to the same level); and the 5-8th and 9-10th items are proportional reasoning items using different-sized containers and balance instruments respectively. After Rasch model analysis, the item and person parameters were estimated, and model fitness was tested by likelihood ratio (Andersen, 1972) and Q1 (van den Wollenberg, 1982) statistics. The test does not satisfy the assumptions of RM (These statistical results were presented in Yang and Hsu, 1989). Here, the graphical presentation (Gustaffsson, 1977) is presented. In figure 2, the predicted and observed probabilities for each ability level are graphed. The first 4 items' patterns are quite different from the last 6 items'. For the first 4 items (especially for the first 3 items), the pattern is over-estimated at low-ability range, under-estimated at middle-ability range, and over-estimated at high-ability range; for items 5-8, the pattern is under-estimated at low-ability range, and over-estimated at high-ability range; for items 9 and 10, the patterns fit the Rasch model much better than the above items. Although we can not explain why, we can hypothesize that the predicted-observed patterns are affected by the latent traits underlying each item and the stems for each item. If this is true, the above analysis might demonstrate the sensitivity of Rasch model.

2. Analysis of TIPS (II)

The second example presented here is the analysis of TIPS (II). Table 2 lists the item-

objective specifications of TIPS (II). There are 9 items for measuring hypothesizing skill. We will present the results of testing whether the 9 items share the same latent trait for different grade students. The details are refered to in Hsu and Yang (1990). By Andersen's Z statistic (1972), it is obvious that these items do not measure the same ability for different grade students. From table 3, the log-likelihood is increasing as grade increases. This is considered as results of the effects of instructions and the properties of each item (Hau and Uang, 1980).

TABLE 3. The log-likelihood of the sub-test of hypothesizing skill

Sample	Log-likelihood
All samples	-5,973.79 (N=1992)
Grade 5	-2,625.86 (N= 717)
Grade 6	-2,318.36 (N= 757)
Grade 7	-937.28 (N= 484)

Z = 92.2584, df = 16, p<.0001

The results of the analysis if TIPS (II) indicate that the items designed to measure "hypothesizing" skill do not underly the same latent trait across different grade testees. By some appropriate techniques. we can detect the bias of test items.

CONCLUSION

The properties of the test models presented in this paper have been well studied in the field of psychological testing, but it is rather new for science educational measurement. However, a test model for educational measurement is not only essential to educational decision making, but also to advance the quality and understand of the nature in our research. Item response theory makes sample-free and test-free measurement, leads to "objective comparable" of different tests and testees. which are very important for achieving scientific science educational researchs.

REFERENCES:

1. Andersen, E.B. (1970). "Asymptotic Properties of Conditional Maximum Likelihood Estima-

tors." *Journal of The Royal Statistical Society* , 32: 283-301.

2. Andersen, E.B. (1972). "The Numerical Solution of A Set of conditional Estimation." *Journal of the Royal Statistical Society* , 34: 42-54.
3. Andersen, E.B. (1973). "A Goodness of Fit Test for the Rasch Model." *Psychometrika* , 38 (1): 123-140.
4. Andersen, E.B. (1973). "Conditional Inference for Multiple-Choice Questionaires." *British Journal of Mathematical and Statistical Psychology* , 26: 31-44.
5. Andersen, E.B. (1980). *Discrete Statistical Models With Social Science Applications* . North-Holland Publishing Company, Amsterdam.
6. Baker, F.B. (1987). "Item Parameter Estimation Under the One-, Two-, and Three-Parameter Logistic Models." *Applied Psychological Measurement* , 11 (2): 111-141.
7. Baker, F.B. (1987). "Item Parameter Estimation Via Minimum Logit Chi-Square." *British Journal of Mathematical and Statistical Psychology* , 40: 50-60.
8. Berka, K. (1983). *Measurement, Its Concepts: Theories and Problems* . D. Reidel Publsh-ing Company, Dordrecht, Holland.
9. Birnbaum. A. (1968). "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability." In Lord, F.M., Novick, M., *Statistical Theories for Mental Test Scores* . Reading, Masschuttets.
10. Burns J.C., Wise, K., Okey , J.R. (1983). "Development of An Integrated Science Process Skill Test." Paper Presented at the Annual Meeting of The National Association for Research in Science Teaching, Dallas.
11. Doran, R. L., (1978). "Measuring the 'Process of Science' Objectives." *Science Education* , 62 (1): 19-33.
12. Embretson (Whitely), S. (1984). "A General latent Trait Model for Response Processes." *Psychometrika* , 49 (2): 175-186.
13. Finley, F.N. (1983). "Science Processes." *Journal of Research in Science Teaching* , 20 (1): 47-54.
14. Fischer, G.H. (1981). "On the Existence and Uniqueness of Maximum Likelihood Estimates in the Rasch Model." *Psychometrika* , 46 (1): 59-77.
15. Fischer, G.H. (1983). "Logistic Latent Trait Models with Linear Contrants." *Psychometrika* , 48 (1): 3-26.
16. Fischer, G.H. (1987). "Applying the Principles of Specific Objectivity and of Generalizability to the Measurement of Change." *Psychometrika* , 52 (4): 565-587.

17. Fischer, G.H., A. Formann. (1982). "Some Applications of Logistic Latent Trait Models with Linear Constraints on the Parameters." *Applied Psychological Measurement* , 6 (4): 397-416.

18. Fischer, G.H., P. Pendl. (1980). "Individualized Testing on the Basis of the Dichotomous Rasch Model." In *Psychometrics for Educational Debates* . Ed. by L.J. Th. van der Kamp, W.F. Langerak and D.N.M.. de Gruijter. John Wiley & Sons Ltd. 171-188.

19. Glas, C.A.W. (1988). "The Rasch Model and Multistage Testing." *Journal of Educational Statistics* , 13 (1): 45-52.

20. Green, K. E., R.M. Smith. (1987). "A Comparison of Two Methods of Decomposing Item Difficulties." *Journal of Educational Statistics* , 12 (4): 369-381.

21. Gustafsson, J.E. (1980). "Testing and Obtaining Fit of Data to the Rasch Model." *British Journal of Mathematical and Statistical Psychology* , 33: 205-233.

22. Gustafsson, J.E. (1980). "A Solution of the Conditional Estimation Problem for Long Tests in the Rasch Model for Dichotomous Items." *Eductional and Psychological Measurement* . 40: 377-385.

23. Gustafsson, J.E. (1980). "An Introduction to Rasch's Measurement Model." Paper presented at the Nordic Researchers' Course Rasch models in the social and behavioral sciences (ED 211 594).

24. Hambleton, R. K. (1983). "Application of Item Response Model to Criterion-Referenced Assessment." *Applied Psychological Measruement* , 7 (1): 33-44.

25. Hambleton, R. K., Swanminathan, J. (1985). *Item Response Theory: Principles and Applications* . Kluwer Nihoff Publishing.

26. Hsu R. F., Yang W.J. (1990). "An Analysis of TIPS (II) by Rasch Model." Paper presented at the Third Sino-Japanese Symposium on Science Education.

27. Hsu, R.F. (1986). "An Emperical Study of Hierarchical Study of Hierarchical Structure in Organizors of Science Process Skills." Paper Presented at the First Sino-Japanese Symosium on Science Education, Taipei, ROC.

28. Hsu, R.F., Yang W.J. (1988). "The Impact of IRT in Measuring the Science Process Skills." Paper Presented in The Fourth Symposium of Science Education, NTNU, ROC. (In Chinese with English Abstract).

29. Jansen, P.G.W., A. L. van der Wollenberg, F.W. Wierda. (1988). "Correcting Unconditional Parameter Estimates in the Rasch Model for Inconsistency." *Applied Psychological Measurement* , 12 (3): 297-306.

30. Lord, F.M. (1986). "An Analysis of the Vearbal Scholastic Aptitude Test Using Birnbaum's Three Parameter Logistic Model." *Educational and Psychological Measurement*, 28: 989-1020.

31. Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Inc.

32. Lord, F.M., Novick, M. (1968). *Statistical Theories of Mental Test Scores*. Reading, Massachusettes.

33. Masters, G.N., Wright, B.D. (1984). "The Essential Process in a Family of Measurement Models." *Psychometrida*, 49 (4): 529-544.

34. Messick, S. (1984). "The Psychology of Educational Measurement." *Journal of Educational Measurement*, 21 (3) 215-237.

35. Okey, J.R., Dillawshaw, F.G. (1979). "Integrated Process Skills Test." Department of Science Education, University of Georgia, Athens, Georgia 30602.

36. Padilla, J.M., Okey, J.R., Dillashaw, F.D. (1983). "The Relationship Between Science Process Skill and Formal Thinking Abilities." *Journal of Research in Science Teaching*, 20 (3) : 239-246.

37. Rasch, G. (1966). "An Item Analysis Which Takes Individual Differences Into Account." *The British Journal of Matematicsal and Statistical Psychology*, 19 (1): 49-57.

38. Rasch, G. (1980, 1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. The University of Chicago Press.

39. Roadrangka, V., Yeany, R.H., and Padilla, M.J. (1983). "The Construction and Validation of Group Assessment of Logical Thinking (GALT)." Paper presented at the annual meeting of the National Association of Research in Science Teaching, Dallas, TX.

40. Spada, H., Mcgaw, B. (1985). "The Assessment of Learning Effects with Linear Logistic Test Models." In Embretson, S.E. (Ed.) *Test Design*. Academic Press, New York.

41. Tatsuoka, K.K. (1987). "Toward an Integration of Item Response Theory and Cognitive Error Diagnoses." This paper is a chapter in the book *Diagnostic Monitoring of Skill and Knowledge* Acquisition, Frederiksen et al., (Ed.); ans is based on a paper presented at the Educational Testing Service Conference. (ED 299 320).

42. Tatsuoka, K.K. (1988). "Indices to Measure Stability of Rule Application." Office of Naval Research, Arlington, Va. Personnel and Training Research Programs Office. (ED 299 319).

43. Tinsley, H.E.A., R.V. Davis. (1977). "Test-Free Person Measurement with the Rasch Simple Logistic Model." *Applied Psychological Measurement*, 1 (4): 483-487.

44. Wainer, H., Morgan, A., Gustafsson, J-E. (1980). "A Review of Estimation Procedures for the Rasch Model With An Eye Toward Lonish Tests." *Journal of Educational Statistics* , 5: 35-64.

45. Warm, T.A. (1978). "A Primer of Item Response Theory." Technical Report 940279, Coast Guard Inst., Oklahoma City, Okla, (ED 171 730).

46. Whitely, S.E. (1980). "Multicomponent Latent Trait Models for Ability Tests." *Psychometrika* , 45 (4): 479-494.

47. Whitely, S. E., Dawis, R.V. (1974). "The Nature of Objectivity With The Rasch Model." *Journal of Educational Measurement* , 11 (2): 163-178.

48. Wood, R. (1978). "Fitting the Rasch model -- A Heady Tale." *British Journal of Mathematical and Statistical Psychology* , 31: 27-32.

49. Wright, B. (1977). "Conditional Versus Unconditional Procedures for Sample-Free Item Analsis." *Educational and Psychological Measurement* , 37: 573-586.

50. Wright, B. (1988). "The Efficacy of Unconditional Maximum Likelihood Bias Correction: Comment on Jansen, van der Wollenberg, and Wierda." *Applied Psychological Measurement* , 12 (3): 315-318.

51. Wright, B.D. (1977). "Misunderstanding of the Rasch model." *Journal of Educational Measurement* , 14 (3): 219-225.

52. Wright, B.D. (1977). "Solving Measurement Problems With the Rasch Model." *Journal of Educational Measurement* , 14 (2): 97-116.

53. Wright, B.D., Douglas, G.A. (1977). "Best Procedures for Sample-Free Item Analysis." *Applied Psychological Measurement* , 1 (2): 281-295.

54. Wright, B.D., M.H. Stone. (1979). *Best Test Design* . MESA Press, Chicago, IL.

55. Wright, B.D., Master, G.N. (1982). *Rating Scale Analysis* . MESA Press, Chicago, IL.

56. Wright, B., Panchapakesan, N. (1969). "A Procedure for Sample-Free Item Analysis." *Educational and Psychological Measurement* , 29: 23-48.

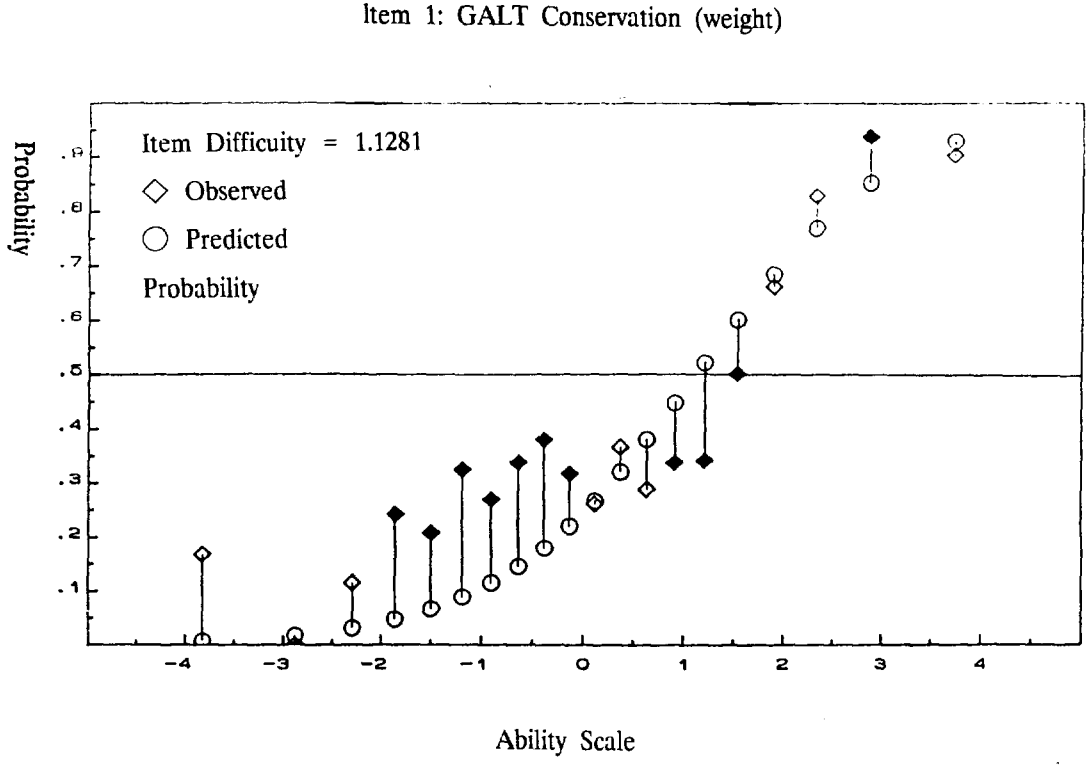
57. van den Wollenberg, A.L. (1982). "Two New Test Statistics for the Rasch Model." *Psychometrika* , 47 (2): 123-140.

58. van der Vijver, F.J.R. (1988). "Systematizing the Item Content in Test Design." In Rolf, L., Rost, J. (Ed.) *Latent Trait and Latent Class Models* . Plenum Press, New York.

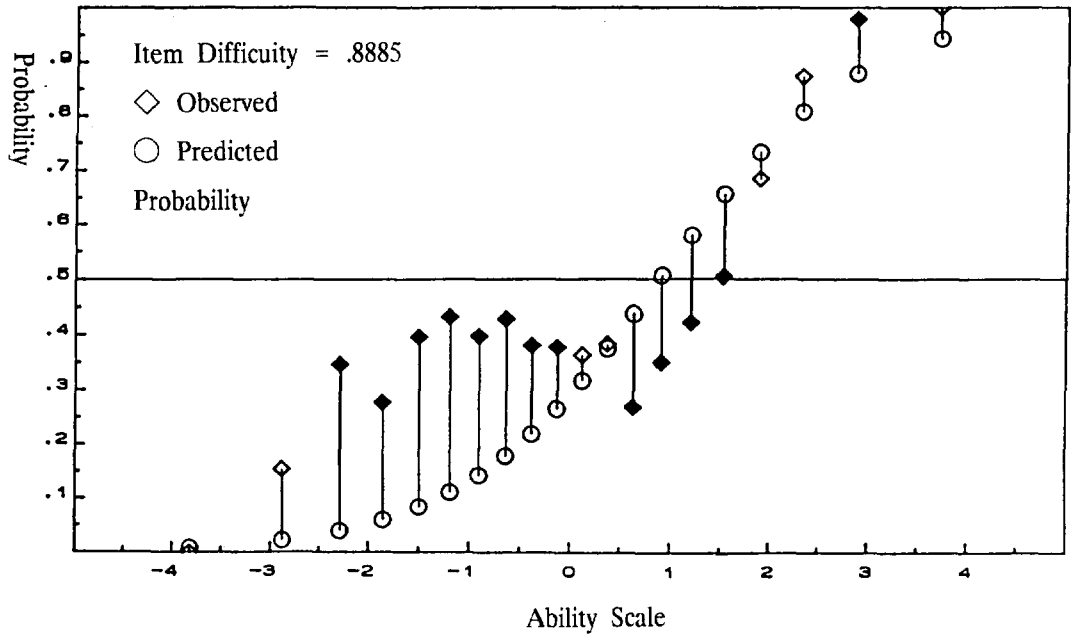
59. van der Wollenberg, A.L., F.W. Wierda, P.G.W. Jansen. (1988). "Consistency of Rasch Model Parameter Estimation: A Simulation Study." *Applied Psychological Measurement* , 12 (3): 307-313.

60. Yang, W.J., Hsu, R.F. (1989). "Item Parameter Estimation of RM--An Empirical Exploration." Paper Pressented in the Fifth Symposium of Science Education, NTNU, ROC.

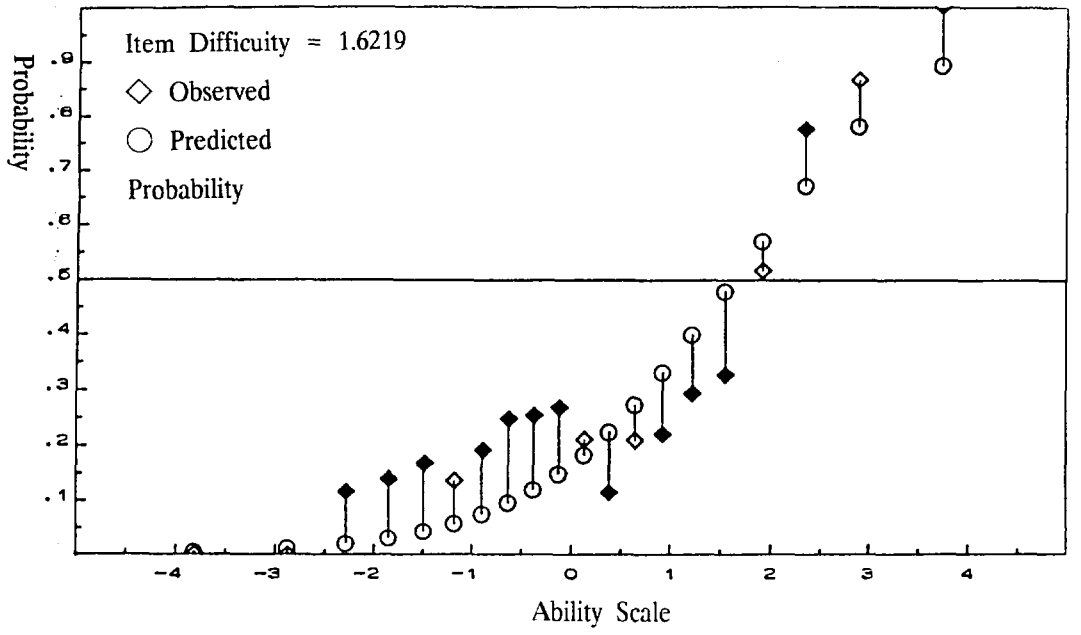
Figrue 2. The predicted-observed probabilities of GALT. The predicted probabilities are expressed as circles, and the connection of circles is the ICC; the observed probabilities are expressed as squares, the solid squares above and under the ICC represent over- and under-estimated cases.



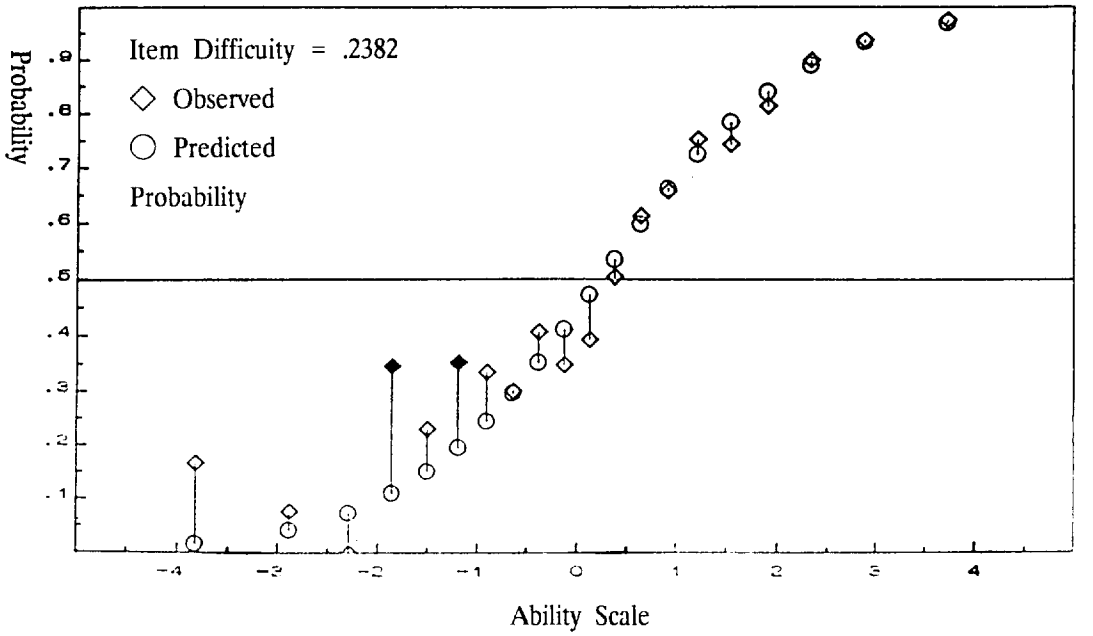
Item 2: GALT Conservation (volume)



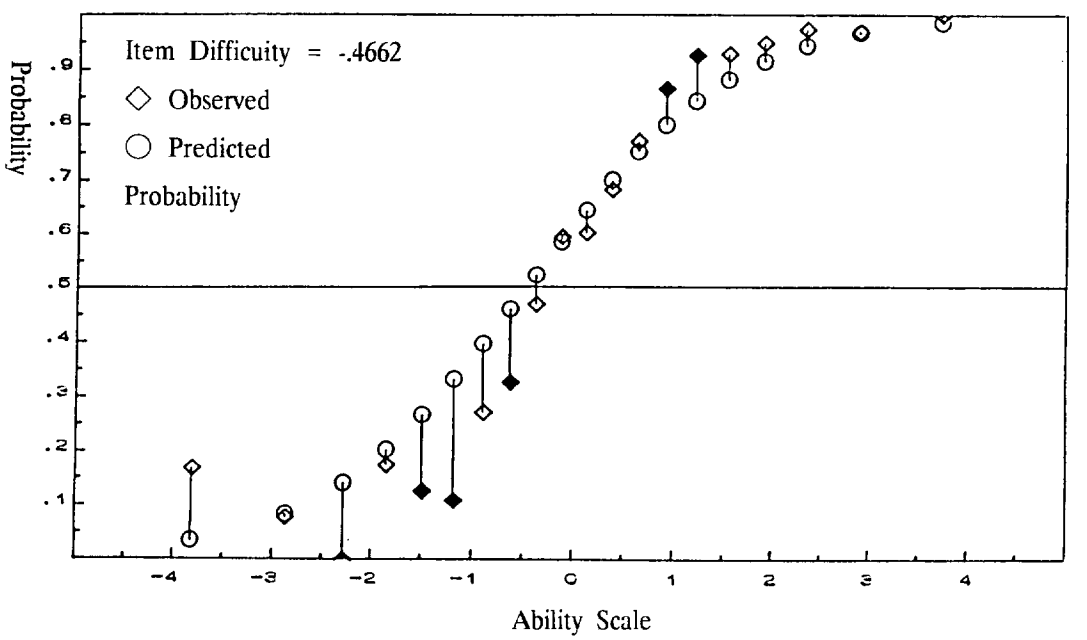
Item 3: GALT Conservation (length)



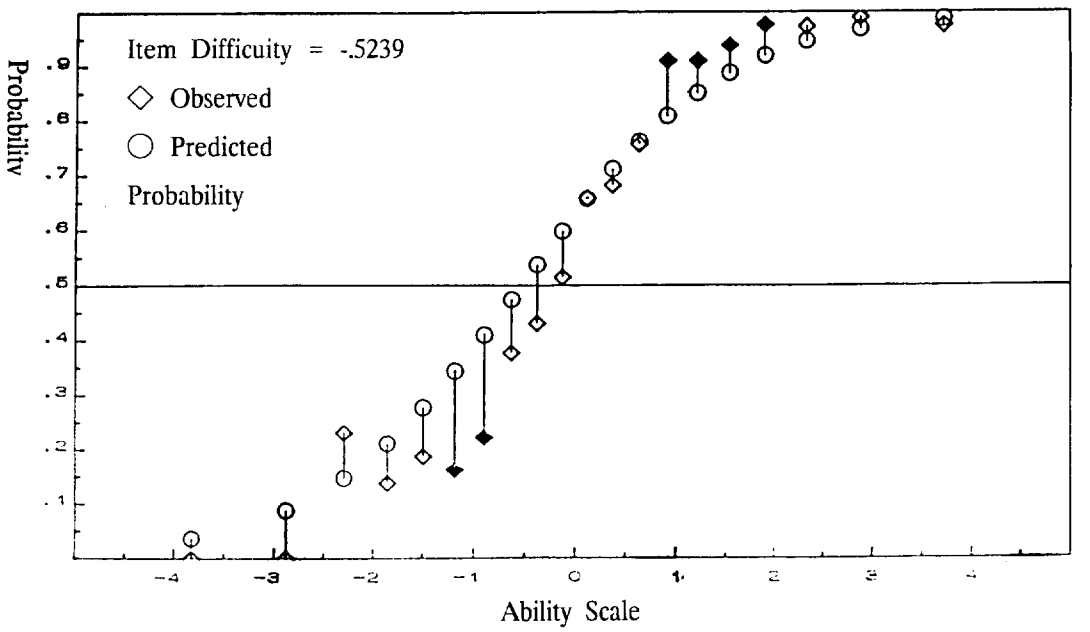
Item 4: GALT Conservation (volume)



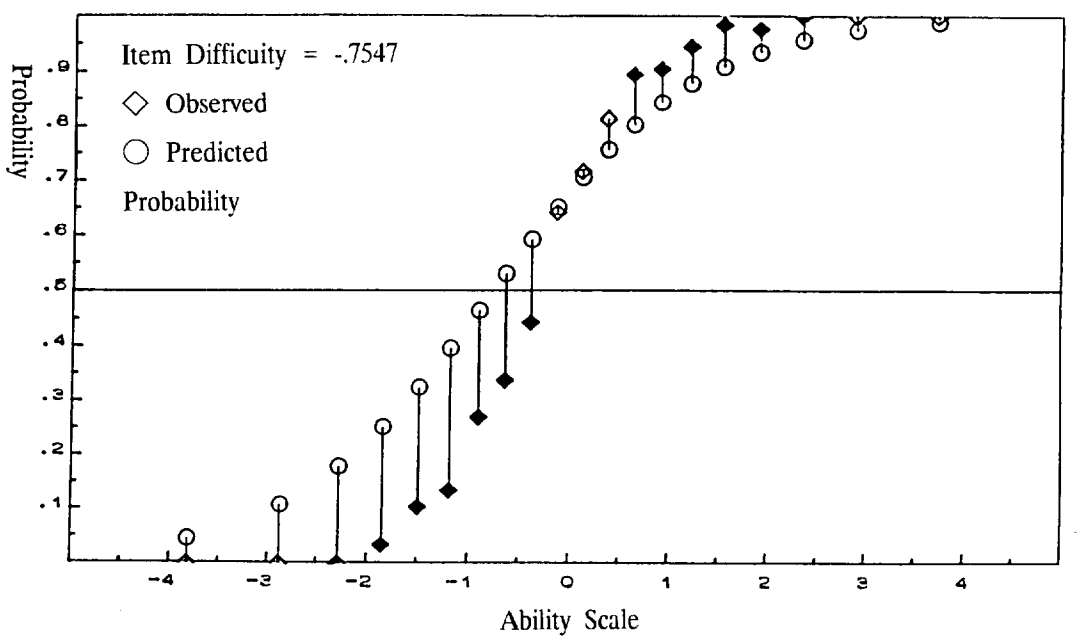
Item 5: GALT Proportion-water



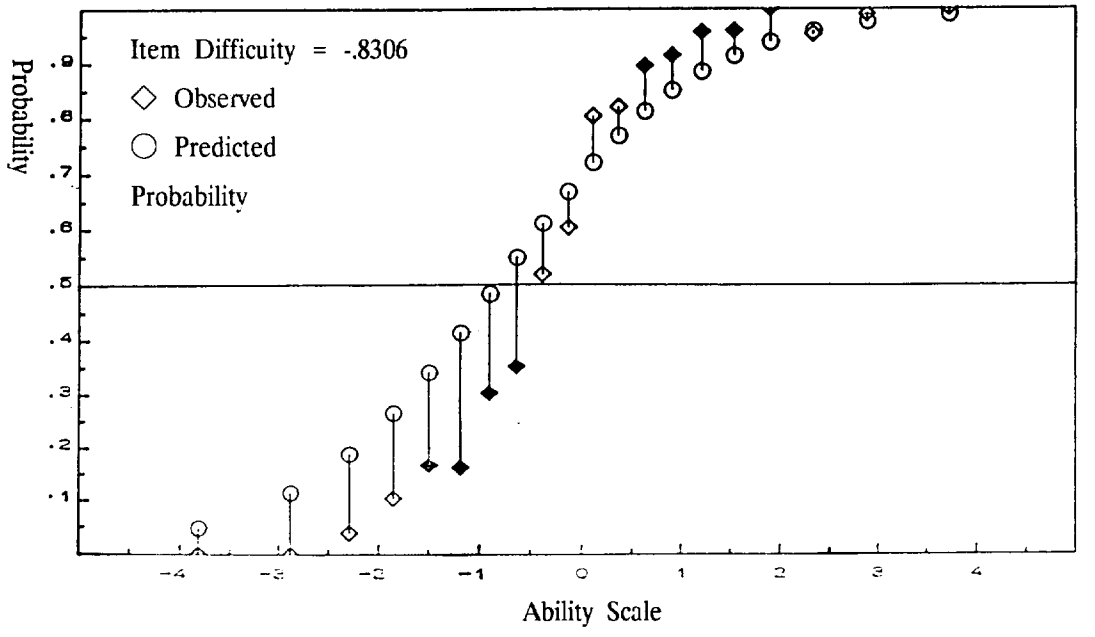
Item 6: GALT Proportion-water



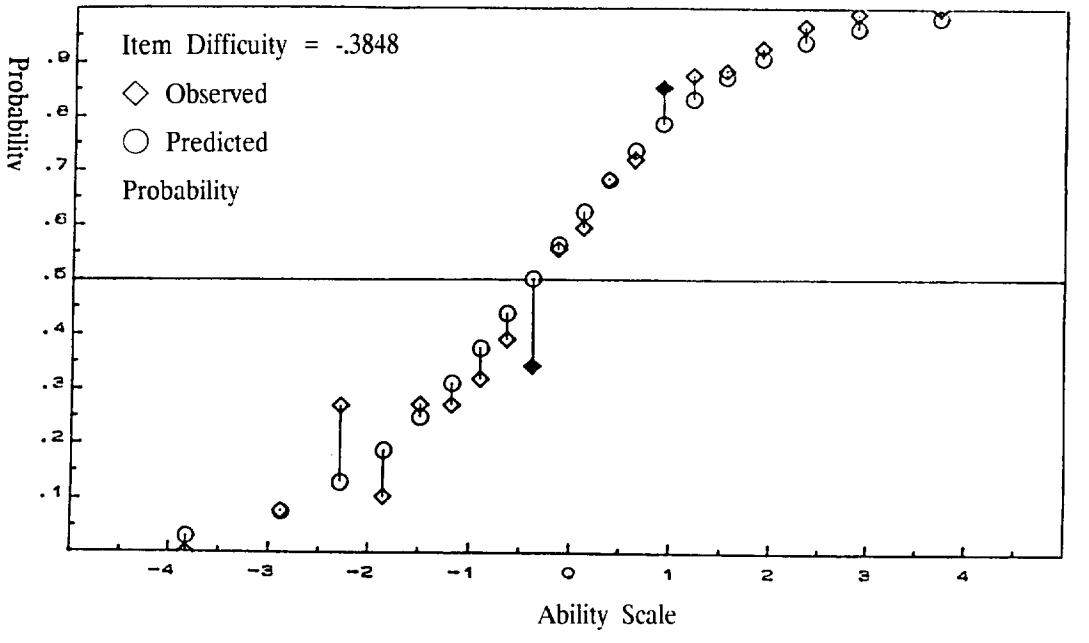
Item 7: GALT Proportion-water



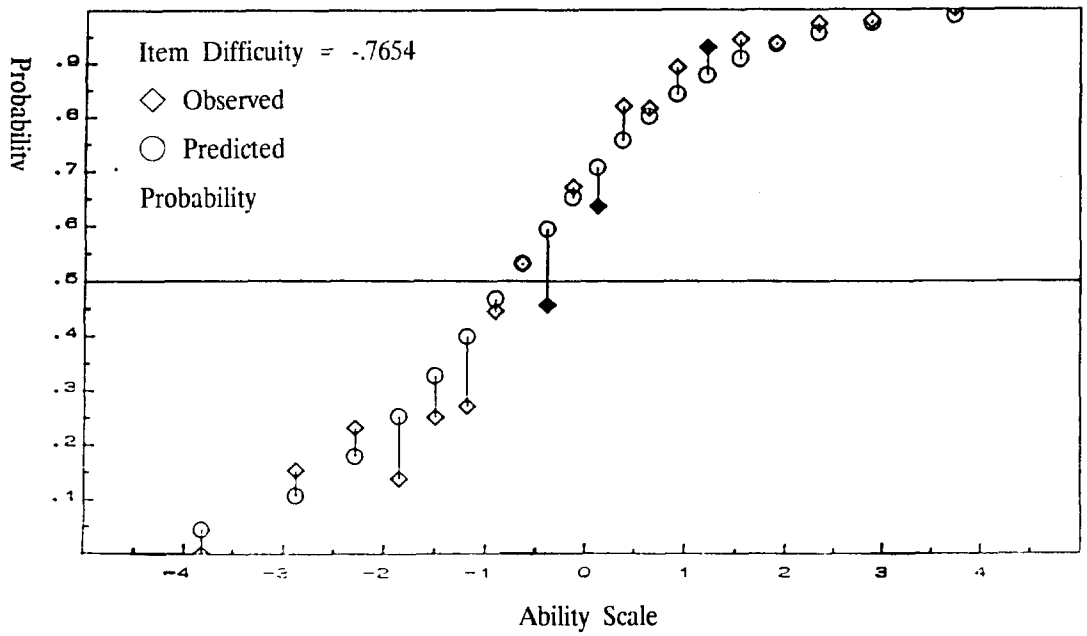
Item 8: GALT Proportion-water



Item 9: GALT Proportion-balance



Item 10: GALT Proportion-balance



統整科學過程技能與邏輯推理能力測驗試題潛在特質分析研究

許榮富

物理系所

摘　要

本文探討一參數試題反應模型（即拉吶模型，Rasch Model, RM）的基本假設及其估計理論，並實際應用於科學教育測驗的分析。

單維性測驗、局部推論獨立性與邏輯型式(logistic form)的試題反應函數爲RM的三個基本假設，據此可得與樣本無關、與試題無關的測量，此特性稱爲「特定客觀性(specific objectivity)」的測量。於本文中，嘗試賦予單維性假設經驗意義。RM的參數估計可有幾種途徑，本文討論非條件與條件最大概度(Unconditional/Conditional Maximum Likelihood)估計的優缺點。除了RM之外，也討論單維的線性邏輯特質模型與多維的成份潛在特質模型等兩個RM的類化模型在科學教育測量上的可應用性。

於此研究中，實際以RM對兩種測驗（TIPS (II)與GALT, 前者爲統整科學過程技能測驗，後者爲邏輯推理能力測驗）進行試題分析，結果顯示理論預測值與觀察值確受試題潛在特質影響，這些影響均以試題特徵曲線分析顯現之。另一方面，結果亦顯示RM在試題特性的偵測上，深具極高鑑別精緻結構及試題偏失的優越性。