We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

18

An Item Response Theory for Peer Assessment

Maomi Ueno
The University of Electro-Communications
Japan

1. Introduction

As Computer Supported Collaborative Learning (CSCL) and other forms of collaborative learning are becoming popular in recent years, peer assessment, i.e., the mutual evaluation among learners, is generating some interest (for instance, Davies, 1999 and Akahori & Kim, 2003). Peer assessment has the following advantages:

1) The learners are more self-reliant and their learning motivation is higher with peer assessment (Weaver & Cotrell, 1986 and Falchikov, 1986).
2) The opinions of other learners are more effective than grade points in inducing the learner's self-reflection (Weaver & Cotrell, 1986).
3) By evaluating others, the assessor is able to learn from the other's work, which induces self-reflection (Falchikov, 1986).
4) Feedback from other learners who have similar backgrounds is readily understood (Falchikov, 1986).
5) It reduces the instructor's workload, and the learner can receive useful feedback even when there is no instructor (Orpen, 1982).
6) Useful feedback which the instructor is unlikely to provide can be obtained; a wide range of feedback can be obtained (Orpen, 1982).
7) When the learners consist of mature adults, evaluation by multiple assessors is more reliable than that by a single instructor (Falchikov, 1986; Orpen, 1982 and Arnold, 1981).

This study is concerned primarily with the advantage 7) above, that is, the use of peer assessment to improve the reliability of evaluations. Falchikov (1986) reports that peer assessments among primary school children were not so reliable, whereas those among junior-high-school students were more reliable. Arnold (1981) introduced peer assessment in a course in medical school, where it was demonstrated that a fair and consistent evaluation took place. Orpen (1982) compared instructor evaluation and peer assessments among university students, and found that not only was there no significant difference between the two when averages were compared, but that peer assessment was in fact more reliable than assessment by a single instructor. The above studies demonstrate that peer assessment is more reliable than an instructor's evaluation, at least in higher education, but there have been no studies so far on methods to further improve the reliability. Furthermore, certain issues remain in peer assessment, such as:

1) The assessors may not all share the same assessment criteria,
2) An assessor may not always be consistent in applying the same assessment criteria,

Fig. 1. The bulletin board and the peer assessment system Title of figure, left justified

3) Treatment of missing data is uncertain.

To resolve these issues, this paper proposes applying Item Response Theory (for instance, Samejima, 1969) to peer assessments, and a method of estimating the parameters. Specifically, we propose a modification of the Graded Item Response model (Samejima, 1969) that incorporates the assessors' evaluation criterion parameters. This model has the following advantages:

1) A consistent assessment based on a common scale is possible even when the assessors have different evaluation criteria.
2) The reliability of the assessors is taken into account to evaluate the learners, which produces a more reliable evaluation.
3) Model parameters can be readily estimated from incomplete or missing data, and the missing data themselves can be estimated.
4) As a result of 1)-3) above, it is possible to assess the learners outcomes with better estimation accuracy.
5) The characteristics of the tasks and assessors can be evaluated from the estimated parameters.

In addition, we propose introducing two indices to analyze assessors: 1) the strictness of the assessor's evaluation criteria, and 2) the assessor's consistency. The proposed method was applied to real data, which demonstrated its validity.

2. Peer Assessment System

The Learning Management System (LMS) "Samurai" (Ueno, 2005), developed by one of the authors, supports a bulletin board system. A learner may set up his/her "room," as shown in Figure 1, where the learner can post his/her assignment work and other remarks.

Students may "visit" the rooms of other students, where they can post a critique of the assignment work, exchange views, and support each other to solve assignment problems (Ueno, 2006). The example of Figure 1 displays a student submitting a weekly report for an undergraduate course on e-learning. The bulletin board, at the lower half of the screenshot, shows the other learners' critiques and opinions of the report. The learner who submitted the report can take these inputs into consideration and rework his/her assignment or rewrite the report. The five buttons shown at the upper left are used for assessing the assignment work, and consist of –2 (Bad), -1 (Poor), 0 (Fair), 1 (Good), and 2 (Excellent). Each room presents an online listing of the average rating and the number of assessors. After converting the points assignment item j, (j=1,…,M) of learner i, (i=1,…,N) given by assessor r, (r=1,…,n), who gave the ranking category x = k, k = 1, 2, …, m (m=5 in the present case), can be obtained as follows: [-2,-1,0,1,2] respectively to [1,2,3,4,5], the data for

$$X = \{x_{ijrk}\}, (i = 1, \cdots, N, j = 1, \cdots, M, r = 1, \cdots, n, k = 1, \cdots, m)$$

where

$$x_{ijrk} = \begin{cases} 1 : \text{when assessor r gives assessment of k to assignment j of learner i} \\ 0 : \text{when otherwise} \bullet \bullet \end{cases}$$

Because all of the element data of X cannot be collected in most cases, each element often contains missing data. This is represented by the missing data of data X as

$$D = \{d_{ijrk}\}, (i = 1, \cdots, N, j = 1, \cdots, M, r = 1, \cdots, n, k = 1, \cdots, m) \quad (1)$$

where

$$d_{ijrk} = \begin{cases} 1 : \text{when there exists an assessment of k by assessor r for assignment j by learner i} \\ 0 : \text{when assessment of k by assessor r for assignment j by learner i does not exist} \bullet \bullet \end{cases} \quad (2)$$

This study proposes applying the item response theory to the data X above.

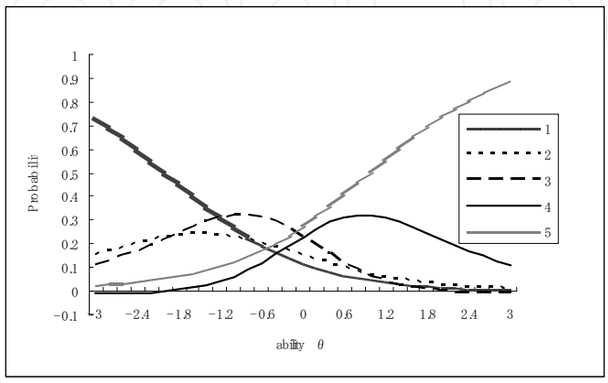


Fig. 2. Examples of response curves for Graded Item Response Model

3. Item Response Theory for Peer Assessment

3.1 Item Response Theory

With the widespread use of computer testing, the Item Response Theory (Samejima, 1969), which is a recent test theory based on mathematical models, is widely being employed in areas such as human resource measurements, entrance exams, and certification tests. It has the following advantages:

1) It is possible to assess ability while minimizing the effects of heterogeneous or aberrant items which have a low estimation accuracy.
2) The learner's response to different items can be assessed on the same scale.
3) Missing data can be readily estimated.

This paper proposes the application of Item Response Theory to data obtained in peer assessments, where the following issues associated with peer assessment can be resolved because of the above advantages.

1) The assessors may not all share the same evaluation criteria,
2) An assessor may not be always consistent in applying the same assessment criteria,
3) Treatment of missing data is uncertain.

While many models have been proposed with regard to Item Response Theory, many peer assessments employ multi-grade Likert Scales (five grades in this study). In this study, therefore, we employ a modification of the Graded Item Response model (Samejima, 1969). This model is used when the assessment of an item can be expressed by points in m grade levels. The probability that the response of learner $i,(i=1,\ldots,N)$ to item $j,(j=1,\ldots,M)$ will be given an assessment of $k,(k = 1, 2, \ldots, m)$ is expressed as follows:

$$p(x_j = k \mid \theta_i) = \frac{1}{1+\exp(-a_j\theta_i + b_{J_{(k-1)}})\}} - \frac{1}{1+\exp(-a_j\theta_i + b_{j_k})} \tag{3}$$

where

a_j: parameter expressing the discriminatory power of item j

b_{jk}: parameter expressing the degree of difficulty of item j with regard to partial point k

θ_i: parameter expressing the ability of learner i

$$p(x_j = 0 \mid \theta_i) = 1 \ ,$$

$$p(x_j = M \mid \theta_i) = 0 \ .$$

Examples of response curves for the model with five grade levels (1-5) are shown in Figure 2, where the abscissa is the learner's ability θ_i, and the ordinate represents the probability that the learner earns grade level (or point) k for this item.Five response curves are shown in the figure, each one corresponding to different grade levels (points). It can be seen that when the learner's ability is low, there is a higher probability of obtaining a lower point, but that when the ability is high, the probability of obtaining higher points increases. We extend this model and propose a modified Item Response Theory for peer assessment.

3.2 Item Response Theory for peer assessment

In this study, the assessors' evaluation criterion parameters are incorporated into the Item Response Theory model. We assume that the response, given in m grade levels, of assessor

r,(r=1,…,n) to assignment work *j,(j=1,…,M)* of learner *i,(i=1,…,N)* is expressed as follows:

$$p(x_j = k \mid \theta_i, \xi_r) = \frac{1}{1+\exp(-a_j\theta_i + b_J + \xi_{r(k-1)})} - \frac{1}{1+\exp(-a_j\theta_i + b_{j_k} + \xi_{rk})} = P^*(k) - P*(k+1) \quad (4)$$

where
a_j: parameter expressing the discriminatory power of assignment *j*
b_j: parameter expressing the degree of difficulty of assignment *j*
ξ_{rk}: parameter expressing strictness of assessor *r*'s evaluation criterion against grade level (point) *k*
θ_i: parameter expressing the ability of learner *i*

For instance, Figure 2 can be viewed as examples of response curves that show assessor *r*'s evaluation of assignment work *j* of learner *i*. In this model, we assume that the characteristics of each evaluation point are determined by the characteristics of assessor *r*. Thus, when ξ_{rk} has a greater value, the assessor's response curve for point *k* will shift to the right, so that it indicates the evaluation criterion strictness of that assessor with respect to point *k*.
The characteristics of assignment *j* are determined only by the discrimination parameter a_j and difficulty parameter b_j. If the discrimination parameter a_j is high, that assignment discriminates well the learner's ability, and the response curves will be spaced apart evenly. If the difficulty paramer b_j is high, all of the response curves will shift toward the right in a parallel manner. In other words, the learner must have a high ability in order to receive a high assessment.

4. Characteristic indices of assessor

We introduce the characteristic indices of assessors which are derived from the estimated assessor parameters.

4.1 Strictness of assessor's evaluation criteria

Denoting by $\hat{\xi}_{rk}$ the estimated assessor parameter, the index b_r , which expresses the strictness of assessor *r*'s evaluation criterion, can be expressed as the average $\hat{\xi}_{rk}$ for all points:

$$b_r = \frac{1}{m-1}\sum_{k=1}^{m-1}\hat{\xi}_{rk} \quad (5)$$

When b_r is large, all response curves will shift to the right, indicating that the learner must have a high ability to receive a high "grade" from this assessor. Conversely, when b_r is

Assignment	a_j	b_j	Subject	Content
#1	1.76 (0.14)	-0.39 (0.04)	The Internet and society	Yesterday, an incident occurred in which a 17-year old youth bashed the head of an infant with a hammer. The arrested youth later testified that seeing various photos of cruelty on the Internet had incited the aggressive impulse. Analyze the causal connection between information access on the Internet and such incidences of violence, and discuss how such incidences may be prevented.
#2	0.61 (0.02)	-0.17 (0.02)	Computers in our lives	Research the extent of information technology use in the local municipality of your hometown, and discuss some of the issues you find.
#3	2.48 (0.18)	0.85 (0.03)	The Internet and privacy	Investigate ways in which private (personal) information can be leaked out via the Internet, and discuss ways of preventing them.

Table 1. Examples of estimated assignment parameters
(figures in parentheses are standard deviations of the estimated parameter values)

Assessor	R_r	b_r	$\xi_{r\cdot\cdot}$	ξ_{r2}	ξ_{r3}	ξ_{r3}
1	0.91	-0.90	-4.02(1.03)	-3.52(0.90)	1.89(0.45)	2.05(0.48)
2	0.71	-1.55	-7.47(0.62)	1.39(0.00)	-1.39(0.00)	1.24(0.42)
3	0.96	0.39	-8.47(2.12)	-5.52(1.25)	6.06(0.36)	9.56(0.89)
4	0.99	3.58	2.29(1.89)	3.18(1.13)	3.98(1.13)	4.87(0.84)
5	0.94	-3.87	-8.29(0.87)	-3.94(1.26)	-2.12(2.08)	-1.22(0.81)

Table 2. Example of estimated parameter values (figures in parentheses are standard deviations of the estimated parameter values)

small, all response curves will shift to the left, and the learner requires only a low ability to receive high "grades."

4.2 Consistency of assessor

It is preferable that the set of parameters ξ_{rk}, which express the strictness of assessor *r*'s evaluation criterion against point *k*, are ordered such that they correspond tothe points *k (k=1,2,…, m)*. For instance,$\xi r2$ should generally be greater than ξ_{r1}. In other words, an assessor whose ξ_{rk} 's are well ordered and evenly spaced can be considered to be consistent. Conversely, an assessor whose parameters display no order whatsoever can be thought to be incongruous and inconsistent. We derive an expression for such an index of consistency.

Thus, the consistency R_r of assessor *r* is defined as the coefficient of correlation between ξrk and *k*, as follows:

$$R_r = \frac{\sum_{k=1}^{m}(\hat{\xi}_{rk} - \bar{\hat{\xi}}_{rk})(k - \sum_{k=1}^{m} k/m)}{\sqrt{\sum_{k=1}^{m}(\hat{\xi}_{rk} - \bar{\hat{\xi}}_{rk})^2}\sqrt{\sum_{k=1}^{m}(k - \sum_{k=1}^{m} k/m)^2}} \tag{6}$$

where $\bar{\hat{\xi}}_{rk} = \frac{1}{m}\sum_{k=1}^{m} \hat{\xi}_{rk}$. The consistency Rr has the maximum value of unity when the values of ξ_{rk} are perfectly well-ordered at equal intervals, zero when they are completely at random, and a minimum value of –1 when they are reversely ordered at equal intervals.

5. Application example

5.1 Data

In this section, we describe an application example of the proposed model using real data. The used data was collected from an e-learning course offered in 2005 on "Information Society and Information Ethics." The details are as follows:

- Initial enrollment: 97 (of which 21 withdrew in midcourse)
- Assignments: submittal of 13 papers, one per week.
- Number of bulletin board comments: 782
- Number of missing data: 384

5.2 Example of estimation of assignment parameters Data

In this section, we present an example of the estimated assignment parameters. Table 1 shows part of the estimated assignment parameters. From the estimated values of the discriminatory power parameter a_j, we see that assignment #2 does not reflect the learners' ability. This assignment required each student to research the extent of information technology employed in his or her hometown (local municipality). It seems that evaluation of the submitted paper was affected more by the actual extent of information technology use in the respective municipalities, rather than the learner's ability, suggesting that this assignment should perhaps be modified. From parameter b_j, representing the degree of difficulty, we see that assignments #1 and #2 were relatively easy, and assignment #3 relatively difficult.

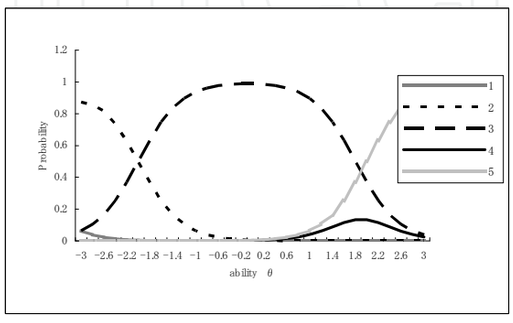


Fig. 3. Response curves of assessor #1

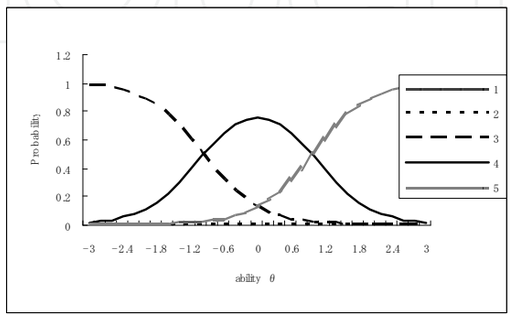


Fig. 4. Response curves of assessor #2

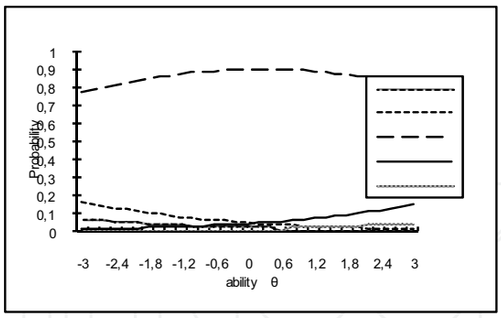


Fig. 5. Response curves of assessor #3

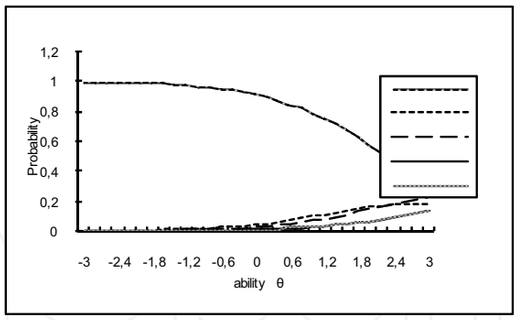


Fig. 6. Response curves of assessor #4

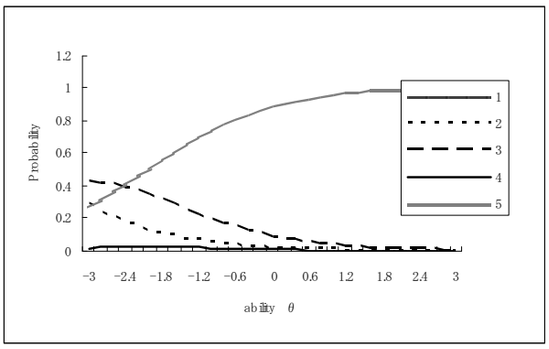


Fig. 7. Response curves of assessor #5

	6-instructor	6-mean	6-θ	13-instructor	13-mean	13-θ
6-instructor	1					
6-mean	0.899	1				
6-θ	0.890	0.782	1			
13-instructor	0.593	0.571	0.511	1		
13-mean	0.875	0.908	0.848	0.548	1	
13-θ	0.848	0.742	0.961	0.510	0.806	1

Table 3. Reliability evaluation of the ability estimators

5.3 Example of estimation of assignment parameters Data

In this section, we present an estimation example of assessor parameters as well as values of assessor reliability. Table 2 shows part of the estimated assessor parameters. The response curves of the assessors are shown in Figures. 3-7, derived from the parameters for the condition that the assignment parameter for discriminatory power is 1 and that for degree of difficulty is zero. Assessor #1 is highly consistent and has appropriate evaluation criteria, but the response curves indicate that he/she has a tendency to give 2, 3, and 5 as the grade point. Assessor #2 has a low consistency, and employs rather lax evaluation criteria. The response curves show that he/she prefers to give 3, 4, and 5 as grades. His/her low consistency is seen by comparing the parameter values, where ξ_{r3} is smaller than ξ_{r2},

suggesting that his/her evaluation criteria are incongruent. Assessor #3 is highly consistent and employs appropriate evaluation criteria, but the response curves in Figure 5 show that he/she has the tendency of giving point 3 in most of his/her evaluations, which means that in effect he/she is not contributing much to the evaluation process. Assessor #4 is highly consistent but employs very strict criteria. Figure 6 shows that he/she gave grade point 1 to most of the learners. Conversely, assessor #5 is highly consistent but is very lax in his/her evaluations. Figure 7 shows that he/she gives point 5 to most learners.

6. Evaluation of estimation accuracy of learner's ability

The greatest advantage of the proposed method lies in the consideration of assessor characteristics, based on which the learners' ability can be expected to be estimated with greater accuracy. In this section, we evaluate the prediction accuracy of the learners' evaluation points when 1) evaluation is done by a single instructor, 2) evaluation is computed from the mean value of peer assessments, and 3) the proposed method is employed. The three types of evaluation values for a total of 13 submitted papers are analyzed by comparing them with those estimated from data on the first six papers. A correlation matrix was computed in which "6-instructor" denotes the evaluation of type 1) based on the first six papers, "6-mean" that of type 2), "6-θ" that of type 3), "13-instructor" the evaluation of type 1) based on all 13 papers, "13-mean" that of type 2), and "13-θ" that of type 3), which is shown in Table 3. The correlation between evaluation values of the assignment data for six papers and 13 papers is highest for the proposed method at 0.961, which indicates its high prediction rate. It is noteworthy that in the evaluation by a single instructor, the correlation between six and 13 papers is extremely low. The observation that the mean value of peer assessments provides a better evaluation than instructor evaluation agrees with the findings of previous studies (Falchikov, 1986; Orpen, 1982 and Arnold, 1981). Furthermore, Ikeda (1992) has reported that when the instructor evaluated an identical student paper twice, with a one-week interval in between, the correlation was still at most around 0.7, suggesting that evaluation by a single instructor is problematic in terms of reliability, especially when there are a large number of learners. In the present case, the instructor's evaluation of six assignments has a correlation over 0.8 with the other methods, but that of 13 assignments has a correlation of around 0.5 with the other methods, showing that the instructor's evaluation diverged considerably from the others. That the present method exhibits a higher reliability than the mean value of peer assessment can be explained by the former's consideration of the heterogeneity existing among the assessors, such as in assessors #2-5 in Table 2. Our findings above show that the evaluation values obtained by the present method have a higher reliability than those obtained from the mean value of the assessors or the instructor's subjective judgment.

7. Conclusion

In this paper, we proposed an application of the Item Response Theory for peer assessment, and discussed a method for parameter estimation. Specifically, The proposed model is a modified Graded Item Response model which incorporates the assessor's assessment criterion parameters. The model was applied to real data, which showed that the present method yields evaluation values that have a higher reliability, with greater predictive

efficiency, than the instructor's assessment or the mean assessment value of all assessors. Our results demonstrate that, in large-scale e learning courses or collaborative learning situations, the application of the present method to peer assessment among the learners yields evaluations that are more reliable than those of a single instructor.

8. References

Davies, P. (1999). Learning through assessment, OLAL on-line assessment and Learning, *Proceedings of the 3rd Computer Assisted Assessment Conference*, pp. 75-78, ISBN: 0-9533210-3-7, Loughborough, UK, June, 1999, The Flexible Learning Initiative, Loughborough University.

Akahori, K. & Kim, S.M. (2003). Peer Evaluation using the Web and some Comparisons Meta-cognition between Experts and Novices, *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA)*, pp. 1484-1487, Honolulu, Hawaii, USA, June, 2003, Chesapeake, VA: AACE.

Weaver, W. & Cotrell, H.W. (1986). Peer evaluation: a case study, *Innovative Higher Education,* Vol. 11, 25-39, ISSN: 0742-5627.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments, *Assessment and Evaluation in Higher Education*, Vol. 11, No.2, 146-166, ISSN: 0260-2938.

Orpen, C. (1982). Student versus lecturer assessment of learning, *Higher Education,* Vol. 11, No.5, 567-572, ISSN: 0018-1560.

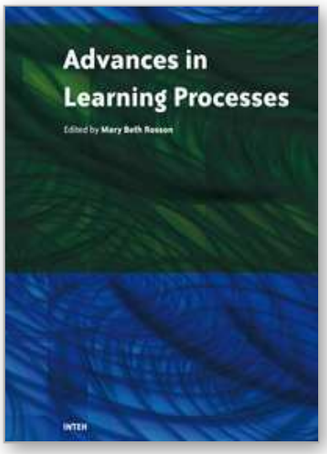
Arnold, L. (1981). Use of peer evaluation in the assessment of medical students, *Journal of Medical Education*, Vol.56, No.1, 35-42.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores, *Psychometric Monograph,* Vol.17.

Ueno, M. (2005). Development of LMS "Samurai" and e-learning practice, *Proceedings of Annual Conference of Educational Information System*, pp. 79-86.

Ueno, M.; Souma, M.; Kinoe, K. & Yamashita, Y. (2006). e-Learning management in Nagaoka University of Technology, *Journal of educational technology*, Vol.29, No.3, 217-229.

H. Ikeda, *Science of test* (Japan culture science publisher, 1992).



Advances in Learning Processes

Edited by Mary Beth Rosson

ISBN 978-953-7619-56-5
Hard cover, 284 pages
Publisher InTech
Published online 01, January, 2010
Published in print edition January, 2010

Readers will find several papers that address high-level issues in the use of technology in education, for example architecture and design frameworks for building online education materials or tools. Several other chapters report novel approaches to intelligent tutors or adaptive systems in educational settings. A number of chapters consider many roles for social computing in education, from simple computer-mediated communication support to more extensive community-building frameworks and tools. Finally, several chapters report state-of-the-art results in tools that can be used to assist educators in critical tasks such as content presentation and grading.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen