A Quantitative Model for Evaluation of Different Forms of Assessment Based on Previous Experiments

M. Mirzarezaee [1], M. Dehghan[2], M. Kharrat[3]
1- Islamic Azad University-Science and Research Branch, Faculty of Engineering
2- Dept. of Computer Eng., Amirkabir University of Technology
3-Iran Telecommunication Research Center (ITRC)
Address: Kareghar Avenue, Tehran, 14399, Iran

Abstract: This paper proposes a quantity model for evaluation of different forms of assessment based on psychometric principle measures obtained from previous experiments. The model consists of five different parameters: *validity*, *reliability*, *fairness*, *comparability* and *learning rate*. It can be used as a decision support tool to help the test administrator on choosing the most suitable form of assessment manually or automatically. The model was tested with results of assessment data collected from a test held in different forms of assessment at some primary schools of Tehran. The obtained results show that learning outcomes are higher when peers are involved.

Key Words: Self-, peer- , collaborative- forms of assessment, Quantity model, Comparative analysis, Collaborative learning and Item Response Theory.

1. Introduction

Recent research in assessment has emphasized on the need to develop forms of assessment with greater relevance to students, and reliability regarding assessment purposes. New forms of assessment have received many attentions in the last decade and several forms of more authentic assessment such as skills of self-, peer- and co-assessment are introduced [1].

With a comparative approach to different forms of assessment, this paper attempts to define a quantity model to express the differences. To do this, we extract parameter values from collected data of previous experiments and ask the test administrator or the tutor on their preferences to finally decide on the best form(s) of assessment.

The structure of this paper is as follows: it starts with an introduction to different possible forms of assessment in a collaborative learning environment. The quantity model and its parameters are shown in section three and finally results of applying the proposed decision model on the collected data are discussed in section four.

2. New forms of assessment

The view that assessment of students’ achievement is something which happens at the end of a learning process is no longer widespread. Assessment is now represented as a tool for learning. The present contribution will focus at one new dimension of assessment innovation, namely the changing place and function of assessor [1].

The current ranges of approaches to assessment are illustrated in Figure 1, in no special order. Figure 1(a) represents a `standard' form of assessment, while diagram 1(b) represents one of the most widely used methods of innovative assessment called self-assessment that refers to the involvement of learners in making judgments about their own learning, particularly about their achievements and outcomes of learning [2].

Figure 1(c) shows another widely used form of assessment, called peer-assessment that is the process whereby groups of individuals rate their peers. The form of assessment often called collaborative assessment is represented in figure 1(d) and is often used in summative assessments. Figure 1(e) represents a form of collaborative assessment, which is called negotiated collaborative assessment. The notion stresses the shared activity typically undertaken by a classroom teacher (or university lecturer) and the student being assessed, to produce an agreed assessment [2].

There are also three other possible ways of assessment shown in figure 1(f) to 1(h), which are a two by two combination of self, peer and collaborative assessment. Self- and peer-assessment are combined when students are assessing peers but the self is also included as a member of the group and must be assessed [1].

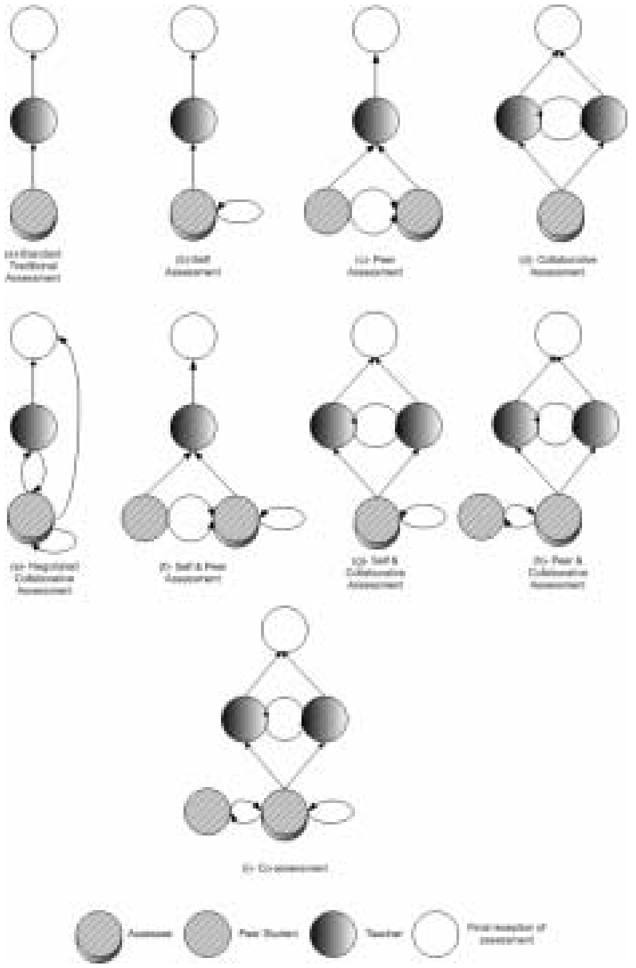
(a)- Standard Traditional Assessment
(b)-Self Assessment
(c)- Peer Assessment
(d)- Collaborative Assessment
(e)- Negotiated Collaborative Assessment
(f)- Self & Peer Assessment
(g)- Self & Collaborative Assessment
(h)- Peer & Collaborative Assessment
(i)- Co-assessment
Assessee
Peer Student
Teacher
Final reception of assessment

Figure 1: Forms of Assessment

Figure 1(i) is a combination of all three forms called co-assessment, which stands for the participation of students with staff in the assessment process; it provides an opportunity for students to assess themselves while allowing the staff to maintain the necessary control over the final assessments [1].

3. A Quantity Model for Evaluation and Analysis of Different Forms of Assessment

We offer a model for the evidentiary arguments that ground differences of forms of assessment, using psychometric principles and show how familiar formulas apply these ideas to familiar forms of assessment, and looks ahead to extending the same principles to new kinds of assessment.

The model provides a way of thinking about psychometrics that relates what we observe to what we infer. It comprises of different parameters, each of which tries to model differences encountered when changing the place and role of the assessor.

To have a good estimation of differences, we assumed that test has been held within the same situations and with equated sets of tasks under one theory of measurement; the only change is the place of the assessor. We used the advantages of Item Response Theory as the underlying method for item parameter estimation and learner assessment.

3.1 Proposed Model Parameters

In practice we do use models, formulas, and statistics to examine the degree to which an assessment argument posses the salutary characteristics of psychometric principles [3].

We do have to consider how these principles are addressed when changing the position of the assessor for a particular purpose. We have proposed five distinct parameters for evaluation of different forms of assessment, called *validity*, *reliability*, *comparability*, *fairness*, and *learning rate*.

These parameters are not just measurement issues, but can also stand for social values that have meaning and force outside the measurement wherever evaluative judgments and decisions are made [3].

3.1.1 Validity parameter

Validity concerns whether the tasks actually do give sound evidence about the knowledge and skills the student model variables are supposed to measure. It speaks directly to the extent to which a claim about a student, based on assessment data from that student is justified [3]. Taking into account just the effect of changing place of the assessor, we defined *validity* measure to be computed as shown in table 1. In this table, *P* stands for Peer, *S* for Self, *C* for Collaborative, *T* for Tutor, and *E* for Expert. N , $N1$, $N2$ are number of assessors of each type (peer, or collaborative).

For computing the precision of validity parameter, a criterion is needed. We used an expert estimation of learner's ability as the criterion value. By expert estimation we mean the estimation of underlying IRT model of learner's ability or simply what an expert human tutor says about learner based on gathered data of his performance during the course or on exams.

The overall value of *validity* parameter for an exam is the average of the *validity* parameter over all students involved in the test and is computed using the following formulas:

$$Validity\_form\_assessment_i = \frac{\sum_{j=1}^{Num\_students} validity_j}{Num\_students} \qquad (1)$$

Table 1 – Validity Parameter Definition

1-Traditional Assessment	$1-\frac{\lvert Ability\_Estimation_E - Ability\_Estimation_T \rvert}{Ability\_Estimation_E}$
2- Self Assessment	$1-\frac{\lvert Ability\_Estimation_E - Ability\_Estimation_S \rvert}{Ability\_Estimation_E}$
3- Peer Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N} Ability\_Estimation_P(i)}{N} \rvert}{Ability\_Estimation_E}$
4-Co Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N} Ability\_Estimation_C(i)}{N} \rvert}{Ability\_Estimation_E}$
5-Self-peer Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N} Ability\_Estimation_P(i) + Ability\_Estimation_S}{N+1} \rvert}{Ability\_Estimation_E}$
6-Self-Co Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N} Ability\_Estimation_C(i) + Ability\_Estimation_S}{N+1} \rvert}{Ability\_Estimation_E}$
7-Peer-Co Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N1} Ability\_Estimation_P(i) + \sum_{i=1}^{N2} Ability\_Estimation_C(i)}{N1+N2} \rvert}{Ability\_Estimation_E}$
8-Self-peer Co Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{\sum_{i=1}^{N1} Ability\_Estimation_P(i) + \sum_{i=1}^{N2} Ability\_Estimation_C(i) + Ability\_Estimation_S}{N1+N2+1} \rvert}{Ability\_Estimation_E}$
9-Nego-Co Assessment	$1-\frac{\lvert Ability\_Estimation_E - \frac{Ability\_Estimation_T + Ability\_Estimation_S}{2} \rvert}{Ability\_Estimation_E}$

3.1.2 Reliability parameter

Reliability concerns the adequacy of data to support a claim. The idea of repeating a measurement process has played a central role in characterizing an assessment's reliability; since variation of the measurements is a good index of the uncertainty associated with that measurement procedure [3].

It is less straight forward to know that what just repeating the measurement procedure means, though, if the procedure has several steps that could each be done differently,(different occasions, tasks, assessors), or if some of the steps can not be repeated at all how much we learn about the student [3,4]. With assumption of just changing the place of assessors, reliability is the probability of similar assessment results from different assessors about the learner's abilities.

A form of assessment is reliable, if variations of the learners' ability estimations, made by different assessors, are less than a pre-specified threshold. Reliability can be defined as follows:

$$reliability = \frac{\sum_{j=1}^{M}\sum_{j=1}^{M}(1-\frac{\lvert Ability\_estimation\_assessor_i - Ability\_estimation\_assessor_j)}{Ability\_estimation\_assessor_j}}{M(M-1)/2} \qquad (2)$$

In above formula, M is the total number of attended assessors in that special form of assessment.

Again the overall value of *reliability* parameter for an exam is computed as the average of that parameter over all students.

3.1.3 Fairness parameter

Fairness is a term that encompasses more territory than we can address. Many of its senses concern social, political, and educational perspectives on the uses to which assessment results inform [3, 4]. *Fairness* depends on some factors: number of assessor, leniency factor, same occasions and equivalent tasks. Here we focus our attention on construct meaning rather than use or consequences, and consider aspect of fairness that bears directly on the place of assessors. In this regard, *fairness* is defined according to the number of assessors involving the assessment. We defined *fairness* as follows:

Results of two different forms of assessment are fair if number of assessors in general and number of tutors, involving their assessment, in particular are the same. Total number of assessors is computed as follows:

$$Number_{Assessors} = Number_{Tutors} + Number_{Peers} + 1 \qquad (3)$$

And *fairness* is computed as shown in table 2. In this table, ε is a very small positive number.

3.1.4 Comparability parameter

Comparability concerns whether what we say about student, base on estimates of his student model variables, has a consistent meaning even if students have taken different tasks, or been assessed under different conditions at different times[3,4]or with different assessors.

As IRT test equating is responsible for equating the estimated abilities obtained from different exams (tasks) [6], we defined *comparability* as follows: the results of two different forms of assessment are comparable if the average expertise of their assessors are the same. The assessors' expertise relates not only on his ability for assessing others but also on his own ability in that special concept.

Therefore expertise and comparability (as we defined it) will change by changing the place of assessor from person to person. In what follows, we proposed a formula for estimation of *comparability* parameter of different forms of assessment based on collected data.

Table 2 – Fairness Parameter Definition

1-Traditional Assessment	$\frac{1}{(Number_{Assessors})^2}$
2-Self Assessment	$\frac{\varepsilon}{(Number_{Assessors})^2}$
3-Peer Assessment	$\frac{\varepsilon * Number_{Peers}}{(Number_{Assessors})^2}$
4-Co Assessment	$(\frac{Number_{Collab}}{Number_{Assessors}})^2$
5-Self-peer Assessment	$\frac{\varepsilon * (number_{Peers} + 1)}{(Number_{Assessors})^2}$
6-Self-Co Assessment	$\frac{(Number_{Collab} + 1)(Number_{Collab})}{(Number_{Assessors})^2}$
7- Peer-Co Assessment	$\frac{(Number_{Collab} + Number_{Peers}) * (Number_{Collab})}{(Number_{Assessors})^2}$
8-Self-peer-Co Assessment	$\frac{Number_{Collab}}{Number_{Assessors}}$
9-Nego-Co Assessment	$\frac{2}{(Number_{Assessors})^2}$

Table 3 – Comparability Parameter Definition

1-Traditional Assessment	$Expertise_{Tutor} = 1$
2-Self Assessment	$Expertise_{Self}$
3-Peer Assessment	$\frac{\sum_{i=1}^{N} Expertise_{Peer}(i)}{N}$
4-Co Assessment	$\frac{\sum_{i=1}^{N} Expertise_{Collab}(i)}{N} = 1$
5-Self-peer Assessment	$\frac{\sum_{i=1}^{N} Expertise_{Peer}(i) + \mathrm{exp}ertise_{Self}}{N+1}$
6-Self-Co Assessment	$\frac{\sum_{i=1}^{N2} Expertise_{Collab}(i) + \mathrm{exp}ertise_{Self}}{N2+1} = \frac{N2 + Expertise_{Self}}{N2+1}$
7-Peer-Co Assessment	$\frac{\sum_{i=1}^{N1} Expertise_{Peer}(i) + \sum_{i=1}^{N2} Expertise_{Collab}(i)}{N1+N2} = \frac{\sum_{i=1}^{N1} Expertise_{Peer}(i) + N2}{N1+N2}$
8-Self-peer- Co Assessment	$\frac{\sum_{i=1}^{N1} Expertise_{Peer}(i) + N2 + Expertise_{Self}}{N1+N2+1}$
9-Nego-Co Assessment	$\frac{1 + Expertise_{Self}}{2}$

Expertise parameter shows how expert our assessor is. We have two different categories of assessors. The first category consists of tutors whom are expert enough in that field of study and also in assessment of their students. The second category is for student assessors. The expertise of self and peers in assessment process is probably less than one because they are not really experts in that field. Their expertise depends on two factors: knowledge of the special course and previous experiments of assessment. We defined expertise as follows:

$$Expertise = \begin{cases} 1 & if \quad assessor \quad is \quad the \quad tutor \quad himself \\ Ability\_assessor * \frac{\sum_{i=1}^{K} i * performance_i}{K} & Otherwise \end{cases} \qquad (4)$$

In the above formula, *k* shows how many times that student attends in the assessment process and $\mathrm{performance}_i$ is the evaluation of his try on assessing classmates based on comparing results with what an expert suggests.

The proposed formulas for estimation of *comparability* parameters are shown in table 3.

3.1.5 Learning Rate

One of the most important parameters to take into consideration is the suitability of the method so that it increases the learning outcomes and students motivations towards learning. A form of assessment is considered to be useful if its learning outcomes are higher than other existing forms of assessment.

Learning rate is computed as follows:

$$\begin{aligned} & Learning\_presicous = 1 - \frac{\lvert Best\_Obtained\_Ability\_score - S \rvert}{Best\_obtained\_Ability\_score} \\ & where \quad S = \frac{1}{N} \sum_{i=1}^{N} Ability\_Estimation\_Student\_by\_Assessor(i) \qquad (5) \\ & where \quad N = number \quad of \quad Assessors \end{aligned}$$

Learning rate of an exam is computed as the average value of the parameter for all students.

3.2 Feedback of the Tutors

Another factor to be considered is the preferences of the test administrator or the tutor to determine the importance of each parameter. The preferences are applied to the final quantity model as the coefficients of each parameter that shows its importance in the final decision making process.

3.3 Quantity Model Formulas

Total evaluation of the system is based on a linear combination of concerned parameters according to the following formula:

$$F_{i+1} = F_i + \gamma_l \sum_{k=1}^{5} \beta_k (previous\_\mathrm{exp}\_parameter_k) \qquad (6)$$

where coefficients (β_k) are the preferences of test administrator (tutor) showing the importance of each parameter on the final decision making process and γ_l determines how much the l-th previous experiment should affect the final decision. The formulas will be calculated separately for each form of assessment. Since the maximum value of the quantity models determines the coordination of forms of assessment with aims and preferences of its test administrator, model can be used for a comparative evaluation of different forms of assessment based on the same preferences.

4 Experimental Results and Conclusions

For model evaluation, the results of a multiplication test held in nine different forms of assessment were used as the test bed. This test was held in eight different primary schools of Tehran, and 120 assessment results were gathered from each school.

For the self-, peer-, collaborative- forms of assessment, we asked the student himself, two classmates and also two teachers to assess his/her performance and inform us about their evaluation of his/her abilities. Table 4 shows the results.

As results shows, peer–collaborative form of assessment obtained the best score for validity and self-collaborative form of assessment, the second grade of comparability. According to the results, fairness of self-peer-collaborative form of assessment is higher among the others. Co-assessment gained the maximum reliability and peer assessment, the maximum learning rate. According to the table, learning rate is higher if peers are involved since the next maximum values are dedicated to the self-peer and self form of assessment respectively; the results confirm the affect of student involvement on their learning outcomes.

Table 4 – Results of applying the quantity model

	Validity Par.	**Reliability Par.**	**Fairness Par.**	**Compara bility**	**Learning Rate**
1 Trad.	0.8237	0.8237	0.040	**1**	0.8436
2 Self	0.7589	0.7589	0.0040	0.7010	0.9105
3 Peer	0.8198	0.9190	0.0080	0.7879	**0.9307**
4 Co	0.8218	**0.9232**	0.1600	**1**	0.8480
5 Self-P	0.8089	0.8595	0.1200	0.7590	0.9239
6 Self-C	0.8272	0.8339	0.2400	**0.9003**	0.8688
7 Peer-C	**0.8465**	0.8575	0.3200	0.8940	0.8893
8 S-P-C	0.8392	0.8383	**0.4000**	0.8554	0.8936
8 Nego.	0.8261	0.7903	0.0800	0.8505	0.8771

In conclusion, we proposed a quantity model for evaluation of different forms of assessment. The model consists of five different parameters: *validity*, *reliability*, *fairness*, *comparability* and *learning rate* which was applied for comparative analysis of different forms of assessment based on previous experiments.

The proposed model can be used as a decision support tool, helping the test administrator on evaluation of applied form of assessment.

The next step is to find parameters related to current assessment situations, and computing their affects on the choice.

References:

[1]- D. Sluijsman, F. Docky, G.Moerkerky, The use of self-,peer- and co-assessment in higher education a review of litreture, *Studies in Higher Education*, Vol. 24, No. 3, p. 331,1999

[2] Paul Brna, John Self, Susan Bull & Helen Pain , Negotiated Collaborative Assessment Through Collaborative Student Modeling, in R. Morales, H. Pain, S. Bull & J. Kay (eds), Proceedings of the Workshop on Open, Interactive and Other Overt Approaches to Learner Modelling, 9th International Conference on Artificial Intelligence in Education, 1999, 35-42.

[3] Mislevy, R.J., Wilson, M.R., Ercikan, K., Chudowsky, N., Psychometric principles in student assessment, In D. Stufflebeam & T. Kellaghan (Eds.), *International Handbook of Educational Evaluation.*, the Netherlands: Kluwer Academic Press, 2002

[4] Robert J. Harvey, Allen L. Hammer, Item Response Theory, Virginia Polytechnic Institute & State University and Consulting Psychologists Press, Inc. harvey.psyc.vt.edu/Documents/TCP_IRT98.pdf, 2000

[5] Embretson, S. E., & Reise, S. P., Item Response Theory for Psychologists. Mahwah, New Jersey: Lawrence Erlbaum Associates, 2000

[6]- Frank B. Baker, *The Basics of Item Response Theor*y, ERIC clearinghouse on Assessment and Evaluation, 2001