

A NEW DIAGNOSTIC ASSESMENT MODEL FOR THE PHYSICS PROBLEM SOLVING PERFORMANCE

Sevket Gunduz^{1,a)}, Mehmet Ali Corlu¹

¹*Marmara University, Faculty of Education Istanbul, Turkey*

1 INTRODUCTION

In this study, a new diagnostic model has been proposed in order to facilitate the analysis of problem solving processes and assessment of problem solving performance of an individual and a group. For the purpose of assessing the performance of a large group, a standard test including multiple choice items should unavoidably be designed [10, 5]. An exemplary test has been prepared in accordance with the proposed model, and it has been implemented to 437 11th grade public high school students, 78 junior students attending public university, and the results have been evaluated.

2 THEORETICAL FRAMEWORK

The main purpose of diagnostic assessment is to identify a person's underlying strengths and needs in a particular area [2, 5]. Such an assessment may be able to explain why a child is experiencing a specific learning difficulty, and can help teachers evaluate the severity of the problem [5, 10, and 18]. In this context, the proposed diagnostic test must give detailed information about the lack of domain contents, level of knowledge application and the failure of reasoning, in addition to identifying the correct results produced by students during the solution of a physics problem. Problem solving consists of a set of logical operations or steps accomplished through certain rules. The results or judgements of each step are obtained with reasoning by using the previously acquired judgements. Therefore, each problem solving step can be defined as the reasoning process. From the rule-based reasoning theories point of view, reasoning can be defined as a process of deriving new information from the knowledge that is explicitly given in accordance with a rule [9, 11, and 19]. A student's success in solving physics problems depends on reasoning processes that can be executed successfully. The diagnostic test must determine the reasoning processes in which students are successful or not. 'The processes and strategies themselves must be the objects of assessment' [10 p: 26]. This is the fundamental assumption of the proposed model. The main structure of the model is based on the measurement of the judgements derived in each step or a reasoning process of the physics problem solving.

Problem solving performance can be evaluated by using different methods such as presenting a problem to students and then *observing* the solving processes [16], designing *interviews* with students [21], and by using special rubrics [8,10]. Interviews can be implemented in alternative ways such as think-aloud, as well as

^{a)} Corresponding author's e-mail: sevketgunduz@gmail.com

fully structured, semi structured, and clinical types. Our model attempts to utilize the advantages of interviews, and eliminate the disadvantages of multiple choice tests. Our new test consists of structured interview questions designed like multiple choice multiple response items. The test has a special construction, as the most difficult and important issue in construction of a test is preparing the items and identifying relations among them.

3 STRATEGY FOR CONSTRUCTING THE DIAGNOSTIC TEST

This strategy defines the main frame of the developed model.

1. A subtopic of physics is selected. For example, "electrostatic force and field".
2. Knowledge elements [14] of the subtopic are determined and stated with suitable sentences with their codes. The knowledge elements can be regarded as content knowledge [4]. A knowledge element represents the scientific definition or principle. These are active in problem solving, and without knowing them it is impossible to carry out the solution [6, 10, and 17].
3. A problem case (context) comprising all concepts of the topic is prepared. Problem case refers to the physical situation that the problem questions are asked about. For example; "Two charged particles hanging vertically by strings in a horizontal electric field are in equilibrium".
4. Inquiring or fundamental questions related to the problem case are produced. For example, "What is the ratio of charge values?"
5. Reasoning map of experts' solutions is prepared by the following methods:
 - a. Fundamental questions are solved by experts in a think-aloud interview [21] and solutions are recorded to analyze.
 - b. Each judgement or knowledge derived in each solution step is stated in separate items and a code number is given to each item.
 - c. A reasoning process representing the derivation of a judgement is defined in statement(s). The reasoning process shows the relation between input and output judgements depending on rules [7, 9, 11, and 19]. There is a causal relation between input and output judgements [7]. This is the hardest step. For example,
"Input judgement: particles are in equilibrium,
output judgement: so resultant forces acted on bodies are zero,
rule: because of the first rule of equilibrium".
 - d. A reasoning map representing knowledge elements, judgements and their relations with each other is prepared. This map also means the graphical representation of reasoning processes with relations (see fig.1).
6. A test item measuring each judgement is prepared by using the reasoning map. Item choices are designed in accordance with the probable responses that students give. These probable responses are gathered by asking the items without choices. Multiple choices multiple response items must be

- preferred in order to assess the given answers with partial crediting and to get the wrong and right answers simultaneously [13].
7. Individual and group assessment criteria are determined in four levels in terms of general test, subtest, processes and questions [3].
 8. In order to verify the techniques improved and make corrections, the test is subjected to pilot administration, and interviews are held with a group of students [5]. During the interview, the students are asked to answer the test questions again.
 9. Reliability and validity study of the test is conducted using the data obtained from the piloting and interviews [5, 12]. After the evaluation of the results, necessary corrections for the test are made. This application is performed repeatedly until the test gets standardized.

4 IMPLEMENTATION

An exemplary diagnostic test in accordance with the proposed model has been developed. In the developing stages 5 teachers were interviewed. The first version of the test consisting of 92 items and four sections was administered to 63 pupils of the 11th grade as a pilot study. Then it was corrected and the number of items was reduced. The second version with 76 items was administered to 374 students of the 11th grade, and 78 first grade students of public university. Students' physics marks in their schools were also used in reliability study. The number of test items was reduced from 76 to 39 according to the results of reliability and validity studies of the second version.

5 FINDINGS

5.1 Findings about Preparation Stages of Diagnostic Test

As a subtopic of physics, "electrostatic force and field" was selected. 21 knowledge elements belonging to this topic were determined and these were grouped in categories of Coulomb force, electric field and equilibrium state. The last one was added after the preparation of problem case. The elements have been coded with two digits, the first shows the category, and the second shows the sequence. For example, "3.i. The force exerted on negatively charged particle in an external electrical field has a direction opposite to field direction". Codes in Figure 1 that are enclosed by circles show the knowledge elements which act as a rule to direct the reasoning process. Then, **the problem case** and **fundamental questions** have been determined as "*Two point charged particles hanging vertically by strings in a horizontal electric field are in equilibrium. A) What is the ratio of charge values, $q_1/q_2 = ?$ b) What are the direction of external electric field and the signs of charges? c) How do the positions change, if the magnitude of external electric field and charges of point materials and distance of materials change?*" The problem was solved by 5 physics teachers in interviews.

48 judgements were found from the teachers' interviews, these are sentences giving information about the problem solution and produced in each solution step as a product of reasoning processes. Judgements have been coded like "p.equ.a",

“p.cou.b”,...symbols. The first character, “p” means of product of process. Second group of characters, “emu”, “cou”, “efi”,...etc. means the theme of judgements out of eight groups. For example “equ” is about equilibrium, “cou” is about Coulomb force, “efi” is about electric field. The third group of characters indicates the sequence of judgements, i.e. a is the first, b is the second, etc. A few judgements were coded with four category, in that case a new character set was added like “x”, “y”, and “z”. For example “p.equ.b.x” from the Figure 1 means that this is the first part of the second judgement produced as an outcome of reasoning process directed by the rules of vector addition. The judgement is “The force F_e due to interaction of electric field is equal to the force F_c due to mutual interaction i.e. Coulomb Law”. The coding symbols were chosen arbitrary.

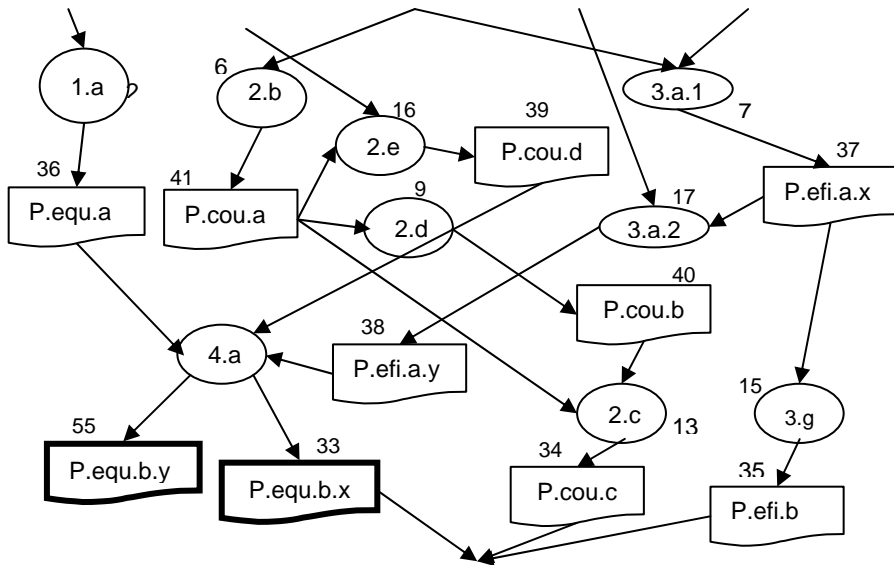


Figure 1 Some parts of the Reasoning Map belong to the second section of diagnostic test of version 2. (The numbers indicate the test question numbers of the second version.)

The reasoning map given in Figure 1 has been prepared after the judgements were determined. The more detailed information about the reasoning maps can be found from the study of Gunduz [8.1]. It shows the reasoning processes used in solution and relations of each process. Each process consists of knowledge elements (enclosed by circles in fig-1) and input-output judgements (enclosed by rectangles). The direction of arrow shows knowledge process from input to output. This map also serves as the objectives that the items measure. Questions measuring the judgements have been prepared accordingly for the diagnostic test. For Example, The 33th question measuring judgement “p.equ.b.x” is that “Which relation is correct between the forces acted on the first particle, F_e due to interaction of electric field and the force F_c due to mutual interaction, in below? (a)

$F_c > F_e$ (a) $F_c = F_e$ (a) $F_c < F_e$ (a) $F_c = 2F_e$ (e) Undefined, because of insufficient data".

5.2 Findings about Students' Reasoning

After the diagnostic test was conducted, the reasoning processes were evaluated and the diagnostic comments have been made for the feedback process. By tracing the responses of questions relative to judgements over the map, the steps that students have failed were determined. Wrong answered questions show the source of failure. Three reasons for the failure in one process have been found out. The first is that the input judgements may not be available from previous steps (processes), the second is the lack of rule in student's memory, and the third is the inability of using the rule to produce the judgement, i.e. inability of knowledge application. As an example, responses of one student who could not solve the judgement "p.sol.b.z" that means "The charges of particles are equal each other" have been analyzed, and the source of failure has been found by tracing backward over the map. The student was unable to remember the knowledge element "3.g" that means "electrical force exerted on charged particle placed in a uniform external electric field is equal $q.E$ ". Evaluation of a single student has been made qualitatively. Item statistics and correlations between the questions were used in quantitative evaluation of group's reasoning processes. The mean of the difficulty indexes of output questions was used as an indicator for the performance of the process. This value was found lower as expected than the mean of input questions, because in addition to input knowledge, application skills are required to obtain the result. Mean differences of input and output questions were evaluated by Cohen-d value [20]. This value was evaluated such that if it is high the application skill is low.

5.3 Findings about Reliability and Validity

Firstly it was verified that the test scores have a normal distribution [1]. Within the frame of validity studies, content and construct validity [3, 8] was controlled qualitatively by 5 experts who are physics teachers having master degrees. Having the test questions prepared by using the reasoning maps almost guarantees that the construct validity of the test is high. Another qualitative study of construct validity was conducted by interviews in order to search whether the test measures the solution processes of 25 students randomly selected from the boarding high school.

Quantitative study of construct validity was conducted in three areas, finding consistency (test-retest) coefficient of test, searching the mean differences between upper and lower groups of the sample, and factor analysis [3,5,12]. Correlation between the scores of the first and second version of the test can be considered as consistency coefficient, and was found as $r = 0,655$ (61, $p < 0,001$). This coefficient can also be interpreted in terms of stability reliability [3]. By means of searching the group differences, it has been found that the exemplary the test discriminates between the upper and lower groups of all three samples significantly in the analysis of independent sample t-test. Factor analysis of the test scores has been

conducted for the three sample groups [5]. The questions composing the factors were approximately the same. These factors were also similar to the group of questions composed at test construction phase. This result shows that construct validity of the test is high.

Findings about criterion related validity were obtained from the high schools and university sample. Correlation coefficient between the test scores and criterion marks which are the average of the physics marks obtained until implementation of the test are $r_{\text{cons}} = 0,44$ (153, $p < 0.01$) for High School, and $r_{\text{cons}} = 0,618$ (78, $p < 0.01$) for the university sample. This is called consistency validity. By means of the predictive validity, correlation coefficients between test scores and the average of the physics marks obtained after the implementation of the test have been calculated as $r_{\text{pred}} = 0,43$ (153, $p < 0.01$), and $r_{\text{pred}} = 0,569$ (78, $p < 0.01$) for both samples. These statistically significant values show that the test has high criterion related validity.

Findings about reliability in terms of internal consistency can be given by the Cronbach-alpha coefficient [12,3]. The results were found as 0,924. Variance analysis can be used in the reliability studies [3]. In this analysis, the ratio of variances coming from between measure and people over residual variance has been searched whether or not it is significant. It has been found that there are statistically significant differences between measure and between people, i.e the F values coming from the two ratios were statistically significant. The differences between people are expected in all tests. In our proposed test the differences between measures are also expected, since an examinee can not succeed in all processes needed to solve the problem.

Qualitatively, high reliability was tested with the interview conducted with 10 students selected randomly from high school. In this interview, it has been searched whether the written and oral answers of the examinee are consistent with each other.

5.4 Findings about Item Analysis

Item analysis has been conducted in five different variables [3, 5]. These are difficulty index, discrimination power index (item-total correlation), item-remainder correlation coefficient, significancy value of mean differences between upper and lower groups (t-values), and standard deviation of items. The items whose difficulty index is out of the range [0,20 - 0,80], and item-total and item-remainder correlation coefficient values, and t-values are not significant are corrected or cancelled.

5.5 Findings About Assessment Results of the Test

By means of group assessment, 9 variables were used. These are # of examinee, % of examinee taken the mark over 45, mean of the test (max = 100), Standard deviation of the test, Difficulty index of the test, Reliability index of the test, skewness and standard error of skewness, kurtosis and s.error of kurtosis, Standard Error of Measurement. Evaluation of these variables in Table 1 gives information about the group properties.

In terms of individual assessment, the test gives the number of correct and incorrect responses, raw point, penalty and net point, the sign of choices selected and interpretation, and the status whether these are True or False for each examinee.

Table 1 Findings of the test in terms of group assessment.

Variables	High Sch.	Univer.
# of examinee	153	78
% of successful examinee	90,85	87,179
mean of the test (max=100)	66,65	61,506
Standard deviation of the test	15,26	12,443
Difficulty index of the test	0,667	0,615
Reliability index of the test	0,924	0,865
skewness	-0,223,	-0,368
and standard error of skewness	0,196	0,272
Kurtosis	-0,723	0,452
and s.error of kurtosis	0,390	0,538
Standard Error of Measurement	4,207	4,572

6 CONCLUSIONS AND SUGGESTIONS

Our model comprises the techniques evaluating the individual and group problem solving performance for the diagnostic purposes. Since one of the main objectives of science and physics education is to develop the students' comprehension, understanding, and scientific inquiry abilities [15], the diagnostic assessment of the problem solving has a growing importance. Briefly, the model depends on producing questions inquiring a single problem context. Those research-based questions are related to each other.

In the group assessment, the information about the general performance, homogeneity and normality of the group can be obtained relatively from difficulty index, standard deviation, kurtosis and skewness values. The success of reasoning processes can be evaluated by using Cohen-d values in which average values of difficulty index and standard deviation of input and output questions are used. As magnitude of the Cohen-d value is increasing, application skill of the process is decreasing.

In the individual assessment, frequency of correct and incorrect answers, raw points, penalty and net points can be used effectively. The mistakes of students in the test can be evaluated by investigating the answers marked as incorrect, because incorrect answers also give information about students' structure of knowledge. By using this test, lack of knowledge and mistakes by students in the process of problem solving can be determined easily. In addition to this information, consistency of answers and success of reasoning processes can be evaluated by

analyzing the answers given to the questions measuring the input and output of the processes.

Besides the fact that the test grade is a reliable indicator about the problem solving performance, incorrect answers have a dominant factor in determining the lack of students' knowledge. The information obtained from the analysis of incorrect answers can be used in the feedback process. This test can also be used in scientific investigations related with problem solving.

We have used reasoning maps in diagnostic assessment in this article. The reasoning maps can be used in other areas of teaching, for example teaching the solving physics problems.

ACKNOWLEDGEMENTS

This study was supported by the Scientific Research Projects Department of Marmara University with the project numbered EĞT-117/081004 and dated 08.10.2004.

REFERENCES

- [1] Akgül, Aziz.(1997). *Tıbbi Araştırmalarda İstatistiksel Analiz Teknikleri: "SPSS Uygulamaları"*. YÖK Matbaası (ISBN:975-96359-0-9), Ankara.
- [2] Bejar, Isaac I.(1984). Educational Diagnostic Assessment, *Journal of Educational Measurement*, V:21, No:2, Summer, 175-189.
- [3] Baykul, Yaşar (2000). *Eğitimde ve Psikolojide Ölçme: Klasik Test Teorisi ve Uygulaması*. ÖSYM Yayınları, ANKARA.
- [4] Bloom, Benjamin S. (1982). *Human Characteristics and School Learning*. (1st pbk ed. Original ed.1976) McGraw-Hill.
- [5] Cohen, L., Manion, L. and Morrison, K. (2000). *Research Methods in Education* (5th Ed.). NY:RoutledgeFalmer.
- [6] Fisher, K.M.(2000). 'Meaningful and Mindful Learning'. In K.M. Fisher, J.H.Wandersee, & D.E.Moody (Eds.), *Mapping Biology Knowledge*, Dordrecht, The Netherlands, Kluwer Academic Publishers, 77-94.
- [7] Forbus, Kneneth D. ve Gentner, D.(1986). Causal Reasoning About Quantities. *Proceedings of the Eighth Annual Meeting of the Cognitive Science Society*, Amherst, MA, August 1986.
- [8] Gipps, Caroline V.(1998). *Beyond Testing: Towards a theory of educational assessment*. (3thEd). The Falmer Press (A member of the Taylor&Francis), London. Washington, D.C.
- [8.1] Gunduz, S (2008). Description of Reasoning Processes Used in Solving Physics Problems: Reasoning Maps. *İetc2008 - International Educational Technologies Conference-2008. Proceeding*.
- [9] Johnson-Laird, philip N. (2000). Reasoning. In *Encyclopedia of Psychology* p.75. PsycBook.
- [10] Kulm, Gerald.(1994).*Mathematics Asessment: What works in the classroom*. Jossey-Bass Publisers, San Francisco.

-
- [11] Kurtz, K.J. , Gentner, D., & Gunn, V. (1999). Reasoning. In D.E. Rumelhart & B.M. Bly (Eds.), *Cognitive science: Handbook of perception and cognition* (2nd ed., pp.145-200). San Diego: Academic Press.
- [12] Lodico, M.G., Spaulding, D.T. and Voegtle, K.H. (2006). *Methods in Educational Research: From Theory to Practice*. CA: Jossey-Bass, A Wiley Imprint
- [13] Ma, Xiaoying, (2004). *An investigation of alternative approaches to scoring multiple response items on a certification exam*. Unpublished doctoral dissertation, University of Massachusetts Amherst, USA.
- [14] Merrill, M. David (1999). Instructional Transaction Theory: Instructional Design Based on Knowledge Objects . In Charles M. Reigeluth (Ed.) *Instructional Design Theories and Models*. Vol.II. NJ: Lawrence Erlbaum Associates.
- [15] National Research Council. 1996. *National Science Education Standards*. <http://www.nap.edu/readingroom/books/nse/> (05.09.2006)
- [16] Ogan-Bekiroğlu, Feral. (2004). *Ne kadar Başarılı? Klasik ve Alternatif Ölçme ve Değerlendirme Yöntemleri: Fizikte Uygulamalar*. Nobel Yayın Dağıtım, Ankara.
- [17] Özçelik, Durmuş Ali. (1981). *Okullarda Ölçme ve Değerlendirme*. ÖSYM-Eğitim Yayınları 3, Ankara.
- [18] Rowntree, Derek, (1996). *Assessing Students: How shall we know them?* (9th Ed) Kogan Page, London; Nichols Publishing Company, New York.
- [19] Simon, Herbert A. (1992). Alternative Representation for Cognition: Search and Reasoning. In (eds) *Cognition: Conceptual and Methodological issues*, p:121, Psybook.
- [20] Sheskin, D. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*. (3rd Ed.). Boca Raton, FL: Chapman & Hall/CRC.
- [21] van Someren, M. W. , Bernard, Y.F. , Sandberg, J.A.C. (1994). *The Think Aloud Method: A practical guide to modelling cognitive processes*. Academic Press: London.