Assessing the Effectiveness of Learning-Teaching Physics Through Traditional Laboratory Experiments

Umapati Pattar[1], Vijay H. Raybagkar[2] and Suresh Garg[3]
[1]Department of Physics, Vijayanagar College, Hospet, Karnataka, India.
[2]Principal, SSR Science, Commerce and Arts College, Silvasa, India.
[3]Professor, School of Sciences, IGNOU, New Delhi, India.

E-mail: umapati.pattar@gmail.com

(Received 2 June 2011; accepted 13 September 2011)

Abstract

Physics laboratory teaching is a part of the curriculum of undergraduate studies. An examination is conducted at the end of the course to evaluate the experimental skills. At present there are no standard diagnostic tests to evaluate the procedural understanding on laboratory teaching. Two tools: Tool-1 and Tool-2, consisting of 42 and 43 objective type questions respectively, were designed as a diagnostic test. The information provided by this test was subjected to *item analysis*. Here we report the results of the item analysis.

Keywords: Assessment, Physics Laboratory.

Resumen

El laboratorio de enseñanza de la Física es parte del currículo de estudios de pregrado. El examen se lleva a cabo al final del curso para evaluar las habilidades experimentales. Actualmente no hay pruebas estándar de diagnóstico para evaluar la comprensión de procedimiento en la enseñanza de laboratorio. Dos herramientas, Herramienta-1 y Herramienta-2, consta de 42 y 43 preguntas de tipos objetivo, respectivamente, fueron diseñados como una prueba diagnóstica. La información proporcionada por esta prueba fue sometida a *análisis de ítems*. Aquí mostramos los resultados de los análisis de ítems.

Palabras clave: Evaluación, Laboratorio de Física.

PACS: 01.40.-d, 01.40.Fk, 01.40.gb **ISSN 1870-9095**

I. INTRODUCTION

Laboratory work has always been an integral component of the Physics curriculum at all levels. Researches in physics education in the last few decades have helped in the evolution of instructional objectives in physics laboratories and there has been a shift towards creating new learning environments to promote meaningful engagement in the learning of physics [1, 2, 3, 4, 5, 6, 7]. Laboratories in such models are primarily based on "discovery" learning. In an undergraduate physics laboratory, a student is expected to make precise measurements, hone investigative skills and discover the interplay between experimentation and fundamental principles underlying physical phenomena. In India, such innovations in the teaching-learning of physics have not impacted much at the undergraduate level. In fact, physics laboratory instruction has all along consisted primarily of performing pre-set repetitive experiments; students are made to go through a prescribed series of steps wherein they are advised to verify certain laws/concepts learnt in theory[8, 9].Such a routine exercise neither promotes scientific investigative skills nor an understanding of the subtle interplay between observation and experimentation. As a result, most students tend to view physics as merely an abstract collection of laws, mathematical equations and textbook problems rather than as a way of understanding and modeling physical phenomena. This situation continues to prevail despite some innovations introduced at the UG level by individual teachers/researchers and a few institutions. This includes the use of home kits to perform simple experiments [10] and micro-processor based (on-line) laboratory with appropriately synergized software [11]. However, these have not led to any major reforms in teaching-learning in the conventional UG Physics laboratories in the country. Recently, Mishra *et al*. [7] carried out factor analysis of distance learners' perceptions and discovered guided approach, student-centered learning and assessment, emphasis on self-learning, use of multimedia and innovative non-conventional teaching strategies, increased student participation and emphasis on problem solving as key determinants for improving the quality of learning in physics laboratory.

In such a scenario, it is important to continuously assess student-learning objectively. Standardized multiple-choice tests have proved useful tools for this purpose. A number of such tests have been developed and are available in literature for different areas of physics. Tests for kinematics [12], force [13], motion [14], dc circuits [15, 16], electricity and magnetism[17], and such other topics have increasingly been used widely by physics instructors to measure some aspects of what students learn in both traditional and reform physics courses. For a dynamic and vibrant subject like physics, there is a need to develop more such instruments in other areas to allow instructors to evaluate the understanding of basic concepts of physics better and to evaluate new teaching endeavors for their feasibility. In this paper, we report the tool, which we developed to Assess the Effectiveness of Physics Teaching through Traditional Laboratory Experiments (ASSEFPT-T-TRALEX). We have also discussed the validity of the test in detail.

A. Research-Based Surveys[1]

A cost-effective way to determine the level of knowledge and understanding of a class is to use a carefully designed *research-based survey*, which is usually a 10 to 30-minute machine-gradable test. It could consist of multiple-choice or short-answer questions, or true-false type. It can be delivered on paper with Scantron™ sheets or on computers. It can be delivered to large numbers of students and the results can be analyzed using computers using spreadsheets or more sophisticated statistical analysis tools. (A *research-based* survey is developed on the basis of qualitative research on hard spots, i.e. student difficulties on particular topics, refined, tested, and validated by detailed observations with a large sample.) Broadly, to achieve good surveys requires the following steps:

- Conduct qualitative research to identify the trends about the thinking process among students.
- Develop a theoretical framework to model the responses of students for that particular topic.
- Develop multiple-choice items to elicit the range of expected possible answers.
- Use the results—including the student selection of wrong answers—to facilitate the design of new instructions, new diagnostic and evaluation tools.
- Use the results to guide qualitative research and further improve the outcomes of survey.

This process places development of evaluation tools firmly in the research-redevelopment cycle of curricular reform. Surveys may focus on a variety of aspects of what students are expected to learn in both the explicit and hidden curriculum. *Content surveys* probe student knowledge of the conceptual bases of particular content areas of physics. *Attitude surveys* probe student thinking about the process and character of learning physics. Over the past two decades, physics education researchers have developed surveys that probe topics from mechanics (the Force Concept Inventory) to the atomic model of matter (the Small Particle Model Assessment Test).

II. DEVELOPMENT OF ASSEFPT-T-TRALEX

To develop ASSEFPT-T-TRALEX, a set of instructional objectives was constructed after an extensive review of literature, including university textbooks and laboratory manuals. (Informal discussions were also held with departmental peers by one of us and their views ascertained.) The questions were selected to test knowledge, understanding and application of concepts in the ratio 3:2:1 respectively. Some test items were selected from Indian books because of their suitability in the use of language used to frame questions. Many other questions were generated on the basis of the experiences of the researchers. The test was designed to assess the comprehensive understanding of laboratory course prescribed by the Gulbarga University Gulbarga, for the First year and Second year B. Sc. Students, rather than probing any specific concept in great detail. Large sample size was used to reduce the magnitude of sampling error [12].

Tool -1 covered the topics on properties of matter, Heat, and sound. It was administered on those students who had completed the Ist year (I- II-semesters) practical course of the Bachelor of Science Program of Gulbarga University. Tool-2, covering AC Circuits, DC Circuits, Electromagnetic induction and Light, was administered on the students who had completed the IInd year (III and IV-semesters) practical course. Research Tool-1, with 42 items was administered on 487 students and Research Tool-2, with 43 items was administered on 535 students from 14 colleges, which are listed in Annexure 1, fall under the jurisdiction of Gulbarga University. A time limit of one hour was set for completion of each test.

III. STATISTICAL EVALUATION OF ASSEFPT-T-TRALEX [12]

Soliciting expert opinion is a standard method of assessing the validity of a test. The term "validity," which is not a statistical construct, refers to the extent to which a test actually measures what it purports to measure. Validity can have several aspects [13]. "Face validity" can be determined by a surface level, common sense reading of an instrument; a test would lack face validity if it tested concepts not related to the subject matter. "Content validity" reflects the coverage of the subject matter—does a test cover enough aspects of a specific topic? Both these aspects of validity are typically assessed by expert consensus, as was done with ASSEFPT-T-TRALEX. Using the data from the sample, we performed four statistical tests: three measures focusing on individual test items (item difficulty index, item discrimination index, item point bi-serial coefficient) and one measure focusing on the test as a

[1] Chapter 5 of Teaching Physics with the Physics suite by E. F. Redish.

whole (Ferguson’s δ). In the following sections, we briefly explain each test and discuss the results of our findings.

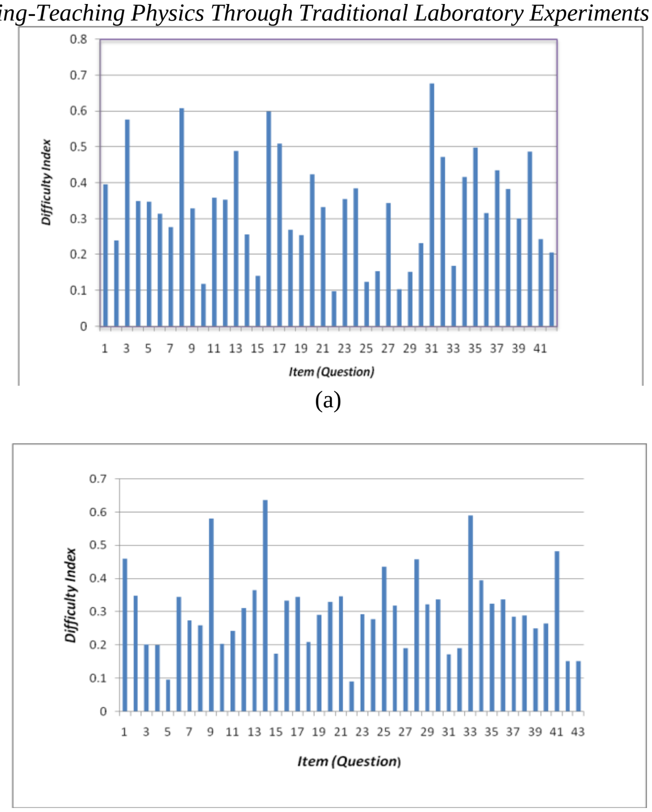
A. Item Difficulty Index

The item difficulty index (*P)* is a measure of the difficulty of a single test question. It is calculated by taking the ratio of the number N_1 of correct responses on the question to the total number *N* of students who attempted the question. Mathematically, we define the difficulty index as

$$P = \frac{N_1}{N}.$$

The difficulty index *P* can also be termed “easiness index,” since it represents the proportion of correct responses on a particular question. The greater is the value of *P*, higher will be the percentage of respondents giving the correct answer and the easier this item is for this population. The range for the difficulty index is [0, 1]. *P=0*, indicates that no one can answer the question correctly, whereas *P*=1 corresponds to the case wherein every one can answer this question correctly. Usually extreme cases should be avoided. A noteworthy aspect of the difficulty index is that the *P- value* depends on the particular population taking the test. There are a number of different possible criteria for acceptable values of the difficulty index for a test [12]. In evaluating ASSEFPT-T-TRALEX, we chose a widely adopted criterion that the difficulty index value lies between 0.3 and 0.9. A difficulty level of 0.5 on each question would lead to the highest values of the statistics discussed in the following sections. However, it is difficult to control every item in one test, especially when the number of items *(K)* in one test becomes large.

The item discrimination index (*D)* is a measure of the discriminatory power of each item in a test. In other words, discrimination index is a measure of the extent to which a single test item distinguishes students who know the material well from those who do not. On a test item with a high discrimination index, students with more robust knowledge will usually answer correctly, while students with weaker understanding will usually get the item wrong. (In contrast, a flawed test question might lead more thoughtful students to give answers that are judged wrong, while students who think less deeply give a correct answer.) If a test contains many items with high discrimination indices, the test itself can be useful in separating intelligent students from weak students in a specific test domain.



(a)

(b)

FIGURE 1. Item difficulty indices for each question for (a) Tool-1 administered on a sample of 487 students, (b) Tool-2 for a sample of 535 students.

An averaged difficulty index value ($\overline{P}$) of all the items in a test gives an indication of the test difficulty:

$$(\overline{P}) = \frac{1}{K}\sum_{i=1}^{K} P_i .$$

We can compare the $\overline{P}$ value with the criterion chosen to check if it meets a certain standard.

Figs. 1(a) & 1(b) show plots of the difficulty index *P* values for each item in ASSEFPT-T-TRALEX for the sample of 487 students [Tool-1] and 535 students [Tool-2]**.** ASSEFPT-T-TRALEX item difficulty index values are 0.34 for Tool-1 and 0.32 for Tool-2.

B. Item Discrimination Index

To calculate the item discrimination index (*D*), we divide the whole sample of students in two groups of equal size, a high group (*H* and a low group (*L*), based on whether an individual total score is higher or lower than the median total score of the entire sample. For a specific test item, one counts the number of correct responses in both groups. Let us denote these as N_H and N_L. If *N* students take the test, the discrimination index *D* of this item can be calculated as

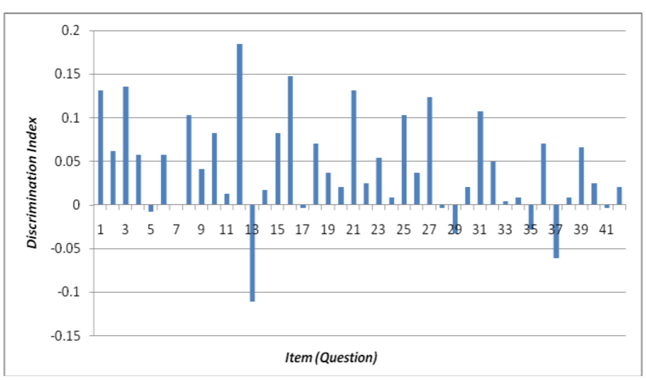
$$D = \frac{N_H - N_L}{N/2}.$$

In educational and psychological studies, there are several different calculations of discrimination index often employed by researchers. The calculation described above (50%–50%) is the one which we adopted to calculate discrimination indices for ASSEFPT-T-TRALEX items. Other researchers may choose to use the top 25% as the high group and the bottom 25% as the low group (25%–25%), in which case the discrimination index D can be expressed as

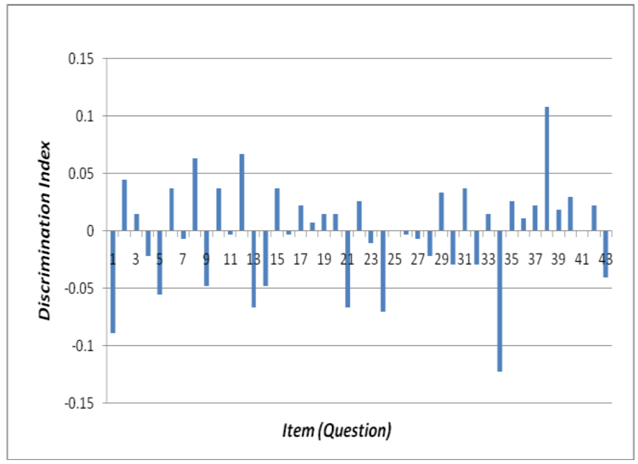
$$D = \frac{N_H(top 25\%) - N_L(bottom 25\%)}{N/4}.$$

The D values obtained on 50%-50% model can underestimate the discriminatory power of test items, since it takes all the students, especially the relatively unstable middle 50%, into account. On the other hand, 25%–25% model uses only the most consistent individuals, reducing the probability of underestimating the discrimination index due to unstable performance, but necessarily discarding half of the available data. The possible range for the item discrimination index D is (−1, +1), where +1 and −1 respectively correspond to the best and the worst values. In the extreme ideal case, all students in the high group get the item correct and all students in the low group get it wrong, giving a discrimination index D of +1. In the worst case, the situation is reversed: everyone in the low group answers correctly, and everyone in the high group gets it wrong. In this case the discrimination index D will be −1.

These extreme cases are unlikely, but it is important to eliminate any items with negative discrimination indices. An item is typically considered to provide good discrimination, if $D \geq 0.3$. Items with a discrimination index lower than 0.3 but greater than 0 are not necessarily bad, but a majority of the items in a test should have relatively high discrimination index values to ensure that the test can distinguish between strong and weak mastery of the material.



(a)



(b)

FIGURE 2. Discrimination indices for each question for (a) Tool-1 for a sample of 487 students, and (b) Tool-2 for a sample of 535 students.

Fig. 2(a, b) shows the discrimination index for each ASSEFPT-T-TRALEX item. We also calculated the averaged discrimination index $\overline{D}$ for all K items (D_i) in ASSEFPT-T-TRALEX. $\overline{D}$ is defined as

$$\overline{D} = \frac{1}{K}\sum_{i=1}^{K} D_i .$$

The average discrimination index $\overline{D}$ for ASSEFPT-T-TRALEX was obtained as 0.04 and -0.001 for Tool-1 and Tool-2, respectively. As may be noted, these values are less than the desired values (refer Table I). There are 8 items (5, 13, 17, 28, 29, 35, 37, 41) in Tool-1 and 19 items (1, 4, 5, 7, 9, 11, 13, 14, 16, 21, 23, 24, 26, 27, 28, 30, 32, 34, 43) in Tool-2 with negative discrimination index. This can be attributed to the Tool having:[2]

- wrong answer;
- more than one correct answer;
- the task is ambiguous;
- the correct choice has a flaw;
- the correct choice is too obvious;
- the task is too difficult and students resort to guessing;
- the item is testing something different from other items;
- the candidates were given wrong information; and
- only the weaker candidates were taught the topic because it was assumed (incorrectly) that better students knew the work.

Careful observation of the items listed above reveals that for meticulously chosen items, the probable flaw could be that the students may not be able to do than that is expected from traditional laboratory course.

[2] Test review after trials.

To illustrate the underestimation of the 50%–50% calculation, we also computed ASSEFPT-T-TRALEX item discrimination indices using 25%–25% model.

C. Point Bi-Serial Coefficient

The point bi-serial coefficient (sometimes referred to as the reliability index for each item) is a measure of the consistency of a single test item with the test as a whole. It reflects the correlation between students' scores on an individual item and their scores on the entire test, and is basically a form of the correlation coefficient. The point bi-serial coefficient has a possible range of [−1, +1]. If an item is highly positively correlated with the whole test, students with high total scores are more likely to answer the item correctly than are students with low total scores. A negative value indicates that students with low overall scores were most likely to get a particular item correct and is an indication that the particular test item is probably defective.

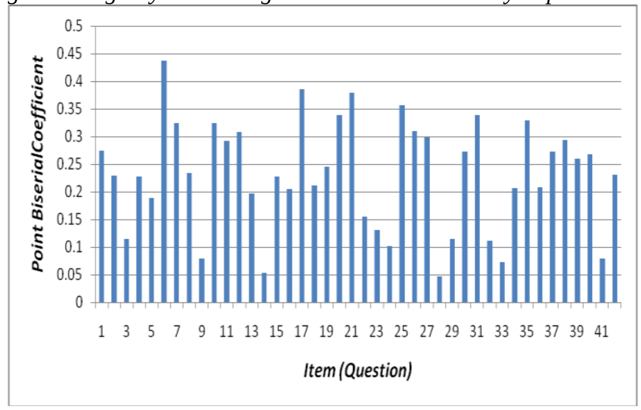
To calculate the point bi-serial coefficient for an item, we calculate the correlation coefficient between the item scores and total scores. A student's score on one item is a dichotomous variable which can have only two values: 1 (correct) or 0 (wrong). Scores for the whole test usually can be viewed as continuous (if the test has a relatively large number of items—say, ≥ 20). The correlation coefficient between a set of dichotomous variables (score for an item) and a set of continuous variables (total scores for the whole test) [12]:

$$r_{pbs} = \frac{\overline{X_1} - \overline{X}}{\sigma_x}\sqrt{\frac{P}{1-P}}\,.$$

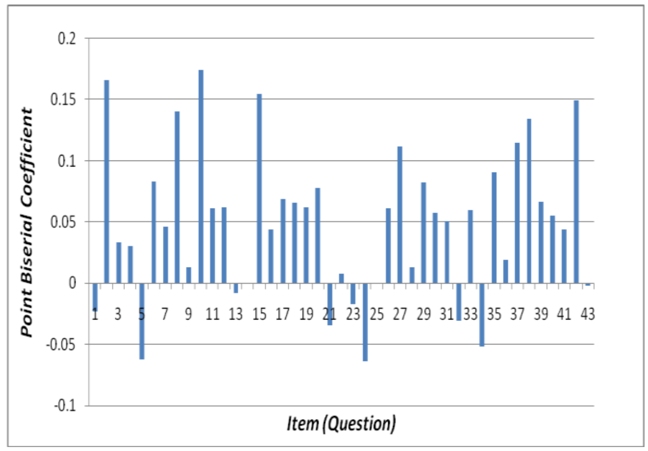
Here $\overline{X_1}$ is the average total score for those students who score 1 for the test item (correctly answer this item), $\overline{X}$ is the average total score for a whole sample, σ_x is the standard deviation of the total score for the whole sample, and P is the difficulty index for this item.

Ideally, all items in a test should be highly correlated with the total score. But this is somewhat unrealistic for a test with a large number of items. The widely adopted criterion [12] for measuring the "consistency" or "reliability" of a test item is $r_{pbs} \geq 0.2$; the average value of r_{pbs} is 0.23 for Tool -1 and 0.05 for Tool -2. This is because 9 items have negative values. Items with point bi-serial coefficient lower than 0.2 can still remain in a test, but there should be a few such items. One way to check whether there are a majority number of items satisfying $r_{pbs} \geq 0.2$ is to calculate the average point bi-serial coefficient of all items (K) in a test:

$$\overline{r_{pbs}} = \frac{1}{K}\sum_{i=1}^{K}\left[(r)_{pbs}\right]_i,$$



3(a)



3(b)

FIGURE 3. Point bi-serial coefficient for each question for (a) Tool-1, and (b) for Tool-2.

where K is the total number of items and $\left[(r)_{pbs}\right]_i$ is the point bi-serial coefficient for the i^{th} - item. Fig. (3a & 3b) provides the point bi-serial coefficient values for each ASSEFPT-T-TRALEX item.

It is worth asking whether or not these two statistics actually measure the same property of an item. The answer is no; theoretically, these two statistics are different measures of an item and could in principle give different results. The item discrimination index is a measure of how powerful an item is in separating strong and weak students, while the point bi-serial coefficient is a measure of whether an item is consistent with the whole test. An item could have a fairly high discrimination index value, but could also show little consistency with the test as a whole. If this were the case, the item might actually be testing some topic that is different from the main subject matter of the rest of the test. On the other hand, an item could be consistent with the test as a whole (high point bi-serial correlation coefficient) but could offer little discriminatory information.

Suppose half of the students in a sample answer a question correctly, giving it an item difficulty index (P) of 0.5. If half of those who answer correctly (25%) have total scores in the top 25% (quartile) and the other half of them (25%) have total scores in the lower mid-25% (quartile), this item would have a fairly high point bi-serial coefficient ($\overline{r_{pbs}}$), but zero discrimination index (*D)* according to the 50%–50% method. This zero discrimination index could be avoided by switching to the 25%–25% discrimination calculation. But this method has its own extreme cases. Suppose only the top 8% of test takers get a particular item correct. Then the point bi-serial coefficient will still be fairly high but the discrimination index (*D* will be lower than 0.3. There are many other possible situations in which the two statistics may be different.

D. Ferguson's Delta

Ferguson's delta is another whole-test statistics. It measures the discriminatory power of a test by investigating how broadly the total scores of a sample are distributed over the possible range. If a test is designed and employed to discriminate among students, one would like to see a broad distribution of total scores. The calculation of Ferguson's delta is based on the relationship between total scores of any two subjects (students). These scores may either be different or equal. If a sample has *N* subjects, the total number of pairs is *N* (*N*-1)/2and the total number of pairs of equal scores is

$$\sum \frac{f_i(f_i-1)}{2} = \frac{\sum f_i^2 - \sum f_i}{2}.$$

Here f_i *represents* the frequency (number of occurrences) of each score. The total number of pairs of different scores is $\frac{\sum f_i^2 - \sum f_i}{2}$. The number of unequal pairs will be greatest if $f_i = N/(K+1)$, where *K* is the number of items. Using this frequency to replace f_i in the above expressions, the number of unequal pairs becomes $\frac{N^2 - \frac{N^2}{K+1}}{2}$, which is the maximum number of unequal pairs a test can provide. The ratio between the number of unequal pairs of scores produced by a test and the maximum number such a test can yield is defined as Ferguson's delta. Accordingly, the expression of Ferguson's delta can be written as

$$\delta = \frac{N^2 - \sum f_i^2}{N^2 - \frac{N^2}{K+1}},$$

where *N* is the number of students in a sample, *K* is the number of test items, and f_i is the frequency (number of occurrence) of cases at each score. The possible range of Ferguson's delta values is [0, 1]. If a test has Ferguson's delta greater than 0.9, the test is considered to offer good discrimination [12].

TABLE I. Summary of statistical results.

Test statistic	Possible values	Desired values	ASSEFT-T- TRALEX value	
			Tool-1	Tool-2
Item difficulty index, P	[0, 1]	≥0.3	Ave 0.34	Ave 0.32
Discrimination index, D	[-1, 1]	≥0.3	Ave 0.044	Ave -0.001
Point biserial Coefficient, r_{pbs}	[-1, 1]	≥0.2	Ave 0.232	Ave 0.05
Frguson' delta	[0, 1]	0.9	0.95	

IV. SUMMARY

The reliability and discriminatory power of the ASSEFPT-T-TRALEX test was evaluated by four statistical tests, three of which focus on individual items and one on the test as a whole. The results, which are summarized in Table -I, indicate that ASSEFPT-T-TRALEX is a reliable test with adequate discriminatory power.

ACKNOWLEDGMENTS

Part of this work was supported by UGC Sanction No. ***MRP(S) – 184/08-09/KAGU028/UGC-SWRO.***

REFERENCES

[1] Mishra, A. V. and Garg, S., *Evaluation of Undergraduate Physics Programme of Indira Gandhi National Open University – A Case Study*, The International Review of Research in Open and Distance Learning **10**(6), 106-123 (2009); http://www.irrodl.org/index.php/irrodl/article/view/753/1456.

[2] Hake, R. R., *Design-Based Research in Physics*, Education Research: A Review, (in Kelly, A. E., Lesh, R. A. and Baek, J. Y. (eds.)), *Handbook of Innovative Design Research in Science, Technology, Engineering, &*

Mathematics (STEM) Education, (Taylor and Francis, USA, 2008).

[3] Heuvelen, A. V. and Et-kina, E., *The Physics Active Learning Guide*, (Pearson Addison Wesley, USA, 2006).

[4] Sokoloff, D. K. and Thornton, R. K., *Inter-active Lecture Demonstrations*, (John Wiley & Sons, USA, 2004).

[5] Laws, P., *Work-shop Physics*, (John Wiley & Sons, USA, 2001).

[6] Jossem, E. L., *Resource Letter EPGA-1: The education of physics graduate assistants*, American Journal of Physics **68**, 502-512 (2000).

[7] McDermott, L. C. and Redish, E. F., *Resource Letter: PER-1: Physics Education Research*, American Journal of Physics **67**(9), 755–767 (1999).

[8] Chakrabarti, B., *The need for the stress on the training in experimental Physics*, IAPT Bulletin **25**, 140 (2008).

[9] Desai, D. A. and Chakrabarti, B., *Experimental Physics in education: need to overhaul the pedagogy*, Physics Teacher **49** (1&2), 89-92 (2007).

[10] Sane, K. V., *Cost-effective science education: the role of educational technology*, Staff and Educational development International **3**, 49-60 (1999).

[11] Jolly, P., *Towards an online Physics laboratory,* Physics Education **19**, 104-110 (1992).

[12] Beichner, R., *Testing student interpretation of kinematicsgraphs*, Am. J. Phys. **62**, 750-762 (1994).

[13] Hestenes, D., Wells, M. and Swackhamer, G., *Force concept inventory*, Phys. Teach. **30**, 141-158 (1992).

[14] Thornton, R. and Sokoloff, D., *Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula*, Am. J. Phys. **66**, 338-352 (1998).

[15] Engelhardt, P. and Beichner, R., *Students' understanding of directcurrent resistive electrical circuits*, Am. J. Phys. **72**, 98-115 (2004).

[16] Engelhardt, P., *Examining students' understanding of electrical circuits through multiple-choice testing and interviews*, Ph.D thesis, North Carolina State University, (1977).

[17] Maloney, D., O'Kuma, T., Hieggelke, C. and Van Heuvelen, A., *Surveying students' conceptual knowledge of electricity andmagnetism*, Am. J. Phys. **69**, S12-S23 (2001).

[18] Best, J. W., *Research in Education, p14,* 4th Ed. (Prentice-Hall, Englewood, 1981).

[19] Lin Ding, R. C., *Physical review special topics - Physics Education research* **2**, 2-6 (2006).

[20] Kline, P., *A Handbook of Test Construction: Introduction to psychometric design*, (Methuen, London, 1986).

[21] Doran, R., *Basic Measurement and Evaluation of Science Instruction*, (NSTA, Washington DC, 1980).

[22] Doran, D., Ref. 13, p.97.

[23] Doran, D., Ref. 12, p. 99.

[24] Ghiselli, E., Campbell, J. and Zedeck, S., *Measurement Theory for the Behavioral Sciences*, (Freeman, San Francisco, 1981).

[25] Kline, P., Ref. 11, p. 143.

[26] Kline, P., Ref. 11, p. 144.