Journal of Technical Science and Technologies; ISSN 2298-0032

An Empirical Study on Assessment of Item-Person Statistics and Reliability Using Classical Test Theory Measurement Methods

Cabir ERGUVEN *
Mehtap ERGUVEN **

Abstract

Recent developments in statistics and psychology supported an increase in quality of measurement in the educational area. The present study empirically examined the characteristics of item and person statistics of mathematics category of School Olympiad Examination (SOE) within the Classical Test Theory (CTT) measurement frameworks. When constructing tests, especially in examining reliability of a test, CTT approach gives opportunity to obtain required targets easily. In this study, various statistics are used to judge the quality of the items. Mainly, item discrimination, item difficulty, item-total inter correlations are found to clarify real aspect of data. Before determining internal consistency, unidimensionality of mathematics examination is analyzed. In this process, Principal Component Analysis is implemented and using item component correlation, behaviors of items are detected from different point of view. Internal consistency is examined by Cronbach's Alpha. Standard error of measurement is calculated to identify confidence interval around an observed score. As a conclusion, reliability of the mathematics part of the SOE is proved by CTT assumptions.

Keywords: Classical test theory, discrimination index, difficulty index, Cronbach-alpha, principal component analysis, reliability, standard error of measurement, confidence interval

Introduction

In the last decades, educational measurement has come to play an increasingly prominent role in education. Educational tests and assessments are used for a wide variety of functions (Linn, 2010). Cognitive tests, various aptitude tests, achievement tests, university entrance examinations or several opinion surveys are all important to differentiate individuals with respect to their abilities and to determine their level among other test takers. This identification can be done sometimes very easily from directly observable behaviors of examinee but sometimes difficultly because of latent psychological traits of individual.

In educational and psychological testing and measurement, the objective is to describe a characteristic of a subject as a numerical score, which represents the quantity of that characteristic of that subject. A quantitative description of these characteristics allows for comparison across subjects, comparison against criterion, and systematic analyses through statistical or other quantitative techniques (Suen, H. K., 1990).

Recently, International Black Sea University organizes School Olympiad Examination (SOE) for the 12th grade students in Tbilisi, Georgia. In this article, among various measurement theories, CTT is selected to analyze the SOE.

The main goal of this approach is to develop reliable psychological instructions and to assess those instructions precisely through the performance of an individual. CTT has served measurement researchers well for many years (Sharkness, J.; De Angelo, L., 2010). This method utilizes traditional item and sample dependent statistics approaches. Although it has some limitations, it is still explicitly popular.

The purpose of the study is using principles of CTT to detect the real aspect of mentioned examination in detail and to implement base assumptions of the theory empirically. Any research based on measurement must be concerned with the accuracy or dependability or, as we usually call it, reliability of measurement (Cronbach, 1951). Measurement theory is concerned with establishing a linkage between mathematics and the elements in the real world that we wish to study (Beckstead, 2013). In this context, this study concentrated to reveal quantitative aspect of the mathematics examination of the SOE to identify its reliability.

Mathematics examination is examined to find the answers to the following questions with respect to CTT assumptions:

1. How well the items of mathematics examination discriminate high ability students and low ability students?
2. What is the difficulty level of each item?
3. How comparable are item discrimination indices with respect to D-discrimination and point biserial correlation?
4. How reliable is the mathematics examination

* Assoc. Prof., Faculty of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia.
E-mail: erguven@ibsu.edu.ge.
** Lecturer, Faculty of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia.
E-mail: merguven@ibsu.edu.ge.

and what is the relation between items and internal consistency? Is there any item which increase or decrease the reliability?

5. What is the confidence interval of an individual with respect to the standard error of measurement?

The term ''reliability'' has been used over the years to refer to two distinct concepts in measurement theory, stability and equivalence (Beckstead, 2013). The ability of educational tests to yield reliable scores that are comparable from year to year, and have acceptable levels of validity, depends on the sophisticated use of psychometric techniques and statistics (Linn, 2010). Here in this study, using appropriate statistical tools, an empirical study of CTT is presented. Therefore, required scientific background is provided to interpret results of the SOE realistically.

Methodology

Dataset consisted of the item scores obtained on School Olympiad Examination, which was organized by International Black Sea University in 2013. Categories of the examination were History, Georgian Language and Literature, English Language, Mathematics and Geometry.

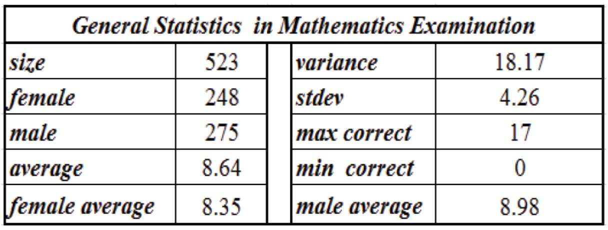
Focus of the research is mathematics category of the SOE and 523 students' test scores are evaluated with respect to CTT principles.

Items were multiple-choice and responses of each student converted into the dichotomous data, correct answer coded as "1" and incorrect answer represented by "0". Microsoft Excel and XLSTAT applications and statistics tools are used to implement classical test-item analyses.

Principal Component Analysis (PCA) is implemented for checking assumption of unidimensionality; representing total variability and illustrating item-component correlations. Graphical illustrations and several tables are represented using MATLAB and XLSTAT.

Using main principles of CTT, difficulty "p", two discrimination indices "D" and point biserial correlation coefficient "rpb" are determined within Excel and MATLAB. Reliability is detected using Cronbach's formula and standard error of measurement (SEM) is defined to describe confidence interval.

General aspect of dataset for the mathematics test is given in table 1. The SOE is administered among 12th grades students, 17 mathematics questions were analyzed for 523 students. 248 female and 275 male students participated to the examination from different regions of Georgia. Average score of mathematics examination is 8.64. This average is 8.35 and 8.98 for female and male students respectively.



General Statistics in Mathematics Examination			
size	523	*variance*	18.17
female	248	*stdev*	4.26
male	275	*max correct*	17
average	8.64	*min correct*	0
female average	8.35	*male average*	8.98

Table1: *General view of data.*

Classical Test Theory

The classical test theory (Gulliksen, 1950) is the earliest theory of measurement. The CTT is referred to as the classical *reliability* theory because its major task is to estimate the reliability of the observed scores of a test. (Suen, H. K., 1990). CTT aims at studying the reliability of a (real-valued) test score variable (measurement, test) that maps a crucial aspect of qualitative or quantitative observations into the set of real numbers (Steyer R. , 1999).

CTT is also regarded as the "true score theory." The theory starts from the assumption that systematic effects between responses of examinees are due only to variation in ability of interest. All other potential sources of variation existing in the testing materials such as external conditions or internal conditions of examinees are assumed either to be *constant* through rigorous standardization or to have an effect that is *nonsystematic* or random (Van der Linden & Hambleton, 2004). The central model of the classical test theory is that observed test scores (OT) are composed of a true score (T) and an error score (E) where the *true* and the error scores are independent. If we were able to administer the test to the same subject under all possible conditions at different times using different possible items, we would have many different observed scores for that subject. The mean of all these *observed scores* would be the most unbiased estimate of the subject's ability. ***This mean is defined as the true score*** (Suen, Principles of Test Theories, 1990)

When the "observed score" of person n on item i is: $x_{ni};\ x_{ni} \in \{0,1,2,\ldots m_i\}$ then the observed total score (e.g., person n's total score on occasion i) under condition of the person on the scale of I items is:

$$OT = \sum_{i=1}^{I} x_{ni} \tag{1}$$

These variables are established by Spearman (1904) and Novick (1966), and best illustrated in the following formula:

$$OT = T + E \tag{2}$$

The classical theory assumes that each individual has a true score, which would be obtained if there were no errors in measurement. However, because measuring instruments are imperfect, the score observed for each person may differ from an individual's true ability (Mango, 2009).

Nevertheless, examinee test scores and corresponding true scores will always depend on the selection of the assessment tasks from the domain of assessment tasks over which their ability scores are defined. Examinees will have lower true scores on difficult tests and higher true scores on easier tests, but their ability scores remain constant over any tests that might built to measure the construct . Of course over time, abilities may change because of instruction and other factors, but at the time of an assessment, each examinee will have an ability score that is defined in relation to the construct and it remains invariant (i.e. independent). (Ronald K. Hambleton, Russel W. Iones).

Discrimination Index of CTT: D

"The correlation between the item score and the total test score has been regarded as an index of item discriminating power" (McDonald, 1999, p. 231). The discriminating power of an item can obviously show the evidence for the quality of the item. About 20 item discrimination indices are proposed. However, based on the previous studies, only a few of them are widely used and compared for the dichotomously scored items.

D is a recognized simpler *discrimination parameter*. First, we need to divide the examinees into the upper and lower groups according to their total test scores. As Kelley (1939) suggested, a more sensitive and stable cut-off point for D is 27% under certain conditions. That means, the top 27% of the examinee group is the *upper group* and the bottom 27% is the *lower group*. Second, we can compute the D through the formula below:

$$D=p_u-p_l \tag{3}$$

Where p_u is the proportion in the upper group who get the item right and p_l is the proportion in the lower group who get the item right.

Ebel (1965) provided the following guidelines based on his own practical experience:

- If $D \geq 0.4$, very well-functioning items.
- If $0.3 \leq D \leq 0.4$ reasonably well functioning items.
- If $0.2 \leq D \leq 0.3$, marginal items which need revised.
- If $D \leq 0.2$, poorly-functioning items which need eliminated or fully revised. (Liu, 2008)

In tests of achievement or ability, negative D value would indicate a poor item in that those who scored most highly on the test overall were not likely to pass the item, whereas those with low overall scores were likely to pass the item. (Kline, 2005)

Point Biserial Correlation

The point-biserial correlation coefficient (rpb) is used to measure the direction and strength of the linear relationship of one factor that is continuous (on an interval or ratio scale of measurement) and a second factor that is dichotomous (on a nominal scale of measurement). The formula for the point biserial correlation is (Privitera, 2011) :

$$r_{pb}=\frac{M_p-M_q}{S_t}\sqrt{pq} \tag{4}$$

Where M_p is the whole-test mean for students answering item correctly, M_q is the whole-test mean for students answering item incorrectly, S_t standard deviation for whole test, p and q are proportion of students answering correctly and incorrectly respectively (Brown, 2001). Important point here is that item score is excluded before the calculation of whole test means M_p and M_q and standard deviation of whole test, for each rpb.

To test for significance, "rpbs" are converted to t-statistics., the t-value formula can be written for "rpb" as (Cohen, 2008):

$$t=\frac{r_{pb}}{\sqrt{\frac{1-r_{pb}^2}{df}}} \tag{5}$$

Difficulty Index of CTT: p

The proportion of individuals who endorse or pass a dichotomous item is termed its p value. In other words, the proportion of examinees passing an item is called difficulty index in CTT. Items with high p values are easy items and those with low p values are difficult items. The variance of a dichotomous item is calculated by multiplying $p \times q$ (where q is the proportion of individuals who failed, or did not endorse, the item). The standard deviation, then, of dichotomous items is simply the square root of $p \times q$. (Kline, 2005)

For dichotomously scored items with respect to previous given model (in calculation of discrimination index), item difficulty, (or p-value) for item j can be defined as:

$$P_j=\frac{t_j}{n} \tag{6}$$

where t_j is the number of correct responses for the item j ; and n is the number of total participants.

Reliability Index: Cronbach Alpha

Both reliability and validity are important in any assessment. Non-reliable and non-valid test scores are simply meaningless numbers (Varma, 2013) .

The alpha formula (Cronbach, 1951) is one of several analyses that may be used to gauge the reliability (i.e., accuracy) of psychological and educational measurements (Cronbach, L.J.; Shavelson, R.J., 2004).

Coefficient alpha can be considered as the lower bound to a theoretical reliability coefficient. The general formula for coefficient alpha is typically written as (Cronbach, 1951):

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_x^2}\right) \tag{7}$$

where k refers to the number of items on the test, σ_i^2 refers to the variance of item i, and σ_x^2 refers to the variance of test scores on the test.

The sum of the item variances should be considered as:

$$\sum_{i=1}^{k}\sigma_i^2 = Var(i_1)+Var(i_2)+\ldots+Var(i_k) \tag{8}$$

Since $cov(i,j)=cov(j,i)$ and from the covariance matrix form as shown below:

$$cov(x)=\begin{bmatrix} \sigma_{i1}^2 & cov(i1,i2) & \ldots & cov(i1,ik) \\ cov(i2,i1) & \sigma_{i2}^2 & & \vdots \\ \vdots & \vdots & & \vdots \\ cov(ik,i1) & \ldots & \ldots & \sigma_{ik}^2 \end{bmatrix}$$

Variance of test scores can be found as:

$$\begin{aligned} \sigma_x^2 &= Var(i1+i2+\ldots+ik) \\ &= Var(i_1)+Var(i_2)+\ldots+Var(i_k)+ \\ &2cov((i_1,i_2)+2Cov(i_1,i_3)+\ldots+2Cov(i_2,i_3)+\ldots \\ &+2Cov(i_k,i_{k-1}) \end{aligned} \tag{9}$$

This given alpha formula reduces to following formula when all items are scored as 1 and zero (Cronbach, 1951):

Kuder and Richardson derive the following formula:

$$r_{tt(KR20)} = \frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}p_i q_i}{\sigma_t^2}\right) \tag{10}$$

where p_i is the proportion receiving a score 1 and qi is the proportion receiving a score of zero on item.

We obtained the same reliability coefficient from both formulas (alpha=0.844).

Classical test theory's reliability coefficients are widely used in behavioral and social research. Each provides an index of measurement consistency ranging from 0 to 1.00 and their interpretation (Webb, N.M.; Shavelson, R.J.;Haertel, E.H., 2006). A value of 0.7-0.8 is an acceptable value for Cronbach's α; values substantially lower indicate an unreliable scale (Field, Reliability Analysis, 2005).

A possible contributor to the confusion is the widespread misunderstanding about the related yet distinct concepts of internal consistency and unidimensionality. Unidimensionality is a subset of consistency. *If a test is unidimensional, it will show internal consistency*. But if a test is internally consistent, it does not necessarily entail one construct (Gardner, 1996).

Hence, it is important to underline the order of implementation steps in determination of reliability process. First, coefficient alpha is useful to estimate reliability in a particular case: when item-specific variance in a unidimensional test is of interest. Second, if among a set of items the existence of unidimensionality is presented, therefore, the computation of Cronbach's alpha for the scale is justifiable and interpretable (Sharkness, J.; De Angelo, L., 2010).

If a test has a large alpha, then it can be concluded that, a large portion of the variance in the test is attributable to general and group factors. This concepts come from Cronbach (1947) and analogous to factor analytic terms (Cortina, 1993).

Determining Unidimensionality with PCA

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002).

Eigen vectors are a set of new basis vectors which transforms "normalized initial data" into the set of principal components PCs. PCA was carried out to obtain the latent root 'Eigen value" from which the principal components were extracted (Erguven, 2012).

Eigen-values are calculated with respect to Eigen vectors using MATLAB. According to the figure 1, first principal component explains 29.36 % of the total variation for all items data with respect to the Eigen values. However, when we compare first Eigen value

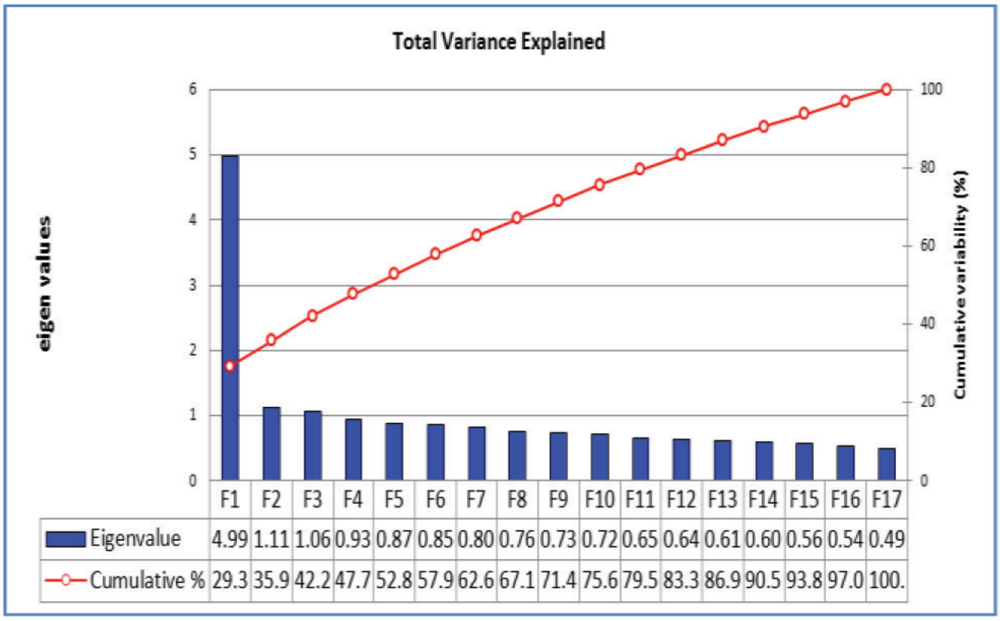


Figure 1

Total Variation for each component (with respect to Eigen values).

with others, remainder Eigen values have significantly low variability (see table 2). This situation is the indicator of unidimensionality and visibly illustrated by the figure 1.

In the SOE Mathematics test, correlations between items and first principle component which is represented as factor loading is satisfactory for all items except item2. Correlation between item2 and PC1 is 0.062 which is very low.

	Factor Loading	
		F1
Variability (%)	item1	0.608
29.362	item2	● 0.062
6.583	item3	0.516
6.275	item4	0.630
5.495	item5	0.571
5.16	item6	0.509
5.026	item7	0.546
4.762	item8	0.589
4.473	item9	0.642
4.312	item10	0.602
4.234	item11	0.635
3.866	item12	0.439
3.782	item13	0.599
3.629	item14	0.521
3.561	item15	0.501
3.349	item16	0.388
3.199	item17	0.569
2.934		

Table 2:

Factor loadings between items and first principal component, and variability of the components with respect to the Eigen values.

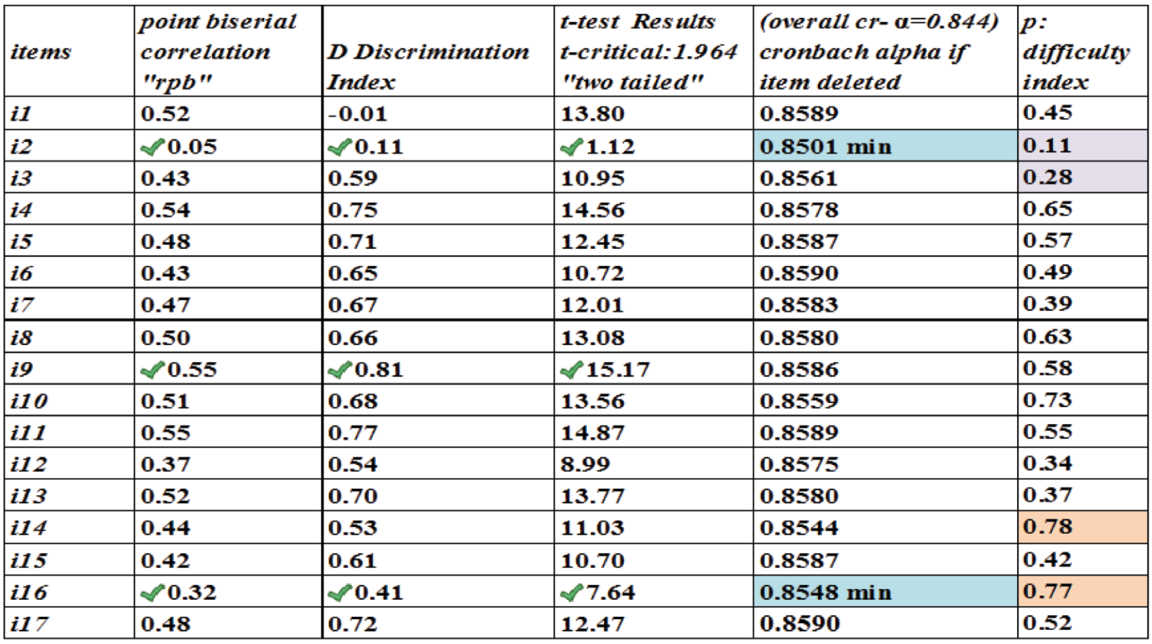
Items showed similar results with respect to the corrected item-total correlations (point biserial correlation coefficients) too.

Item-Person Statistics

In this part items with respect to their discrimination and difficulty indices, and item-total correlations detected one by one. Results are presented and interpreted here in detail:

Higher item-test correlation is desired, which indicates that high ability examinees tend to get the item correct and low ability examinees tend to get the item incorrect (Ji Zeng,Adam Wyse, 2009). Item-test correlation is detected with point biserial coefficients. The strength of associations is moderately high almost for all point biserial coefficients as shown in table3. Item-total point biserial correlation coefficients are considerably less than others for item 12 (rpb=0.37) and item16 (rpb=0.32). Our null hypothesis is, high ability group students and low ability group students have same mean, and there is no significant difference between them (H_0:μ_h=μ_l). Interpretation of rpbs is changeable with respect to number of degrees of freedom and hence for better analysis t-statistics are calculated. In order to represent more accurate results, t-statistics computation connected with rpb values.

According to t-test (see table3), for item12 and item16, calculated t-values are 8.99 and 7.64 respectively. This t-values are still greater than t=1.964 for two tailed t-test (at the level of significance p<0.05). However, "item-total" point biserial correlation of item2 (0.05) is very low and its t-statistic 1.12 is less than

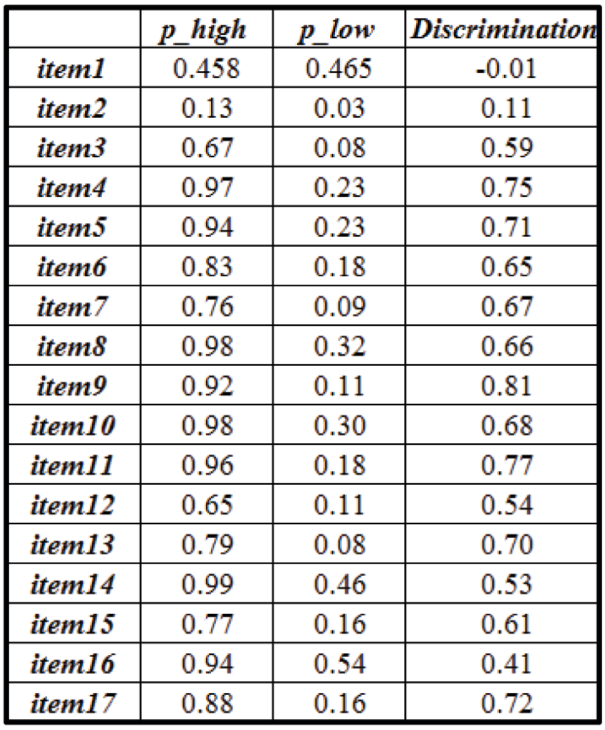


items	point biserial correlation "rpb"	D Discrimination Index	t-test Results t-critical:1.964 "two tailed"	(overall cr- α=0.844) cronbach alpha if item deleted	p: difficulty index
i1	0.52	-0.01	13.80	0.8589	0.45
i2	✓0.05	✓0.11	✓1.12	0.8501 min	0.11
i3	0.43	0.59	10.95	0.8561	0.28
i4	0.54	0.75	14.56	0.8578	0.65
i5	0.48	0.71	12.45	0.8587	0.57
i6	0.43	0.65	10.72	0.8590	0.49
i7	0.47	0.67	12.01	0.8583	0.39
i8	0.50	0.66	13.08	0.8580	0.63
i9	✓0.55	✓0.81	✓15.17	0.8586	0.58
i10	0.51	0.68	13.56	0.8559	0.73
i11	0.55	0.77	14.87	0.8589	0.55
i12	0.37	0.54	8.99	0.8575	0.34
i13	0.52	0.70	13.77	0.8580	0.37
i14	0.44	0.53	11.03	0.8544	0.78
i15	0.42	0.61	10.70	0.8587	0.42
i16	✓0.32	✓0.41	✓7.64	0.8548 min	0.77
i17	0.48	0.72	12.47	0.8590	0.52

Table 3
rpb-point biserial correlation, D-discrimination, p-difficulty, Crounbach alphat, t-test results.

the t-critical value 1.964 (alpha=0.05 for the degrees of freedom 521). Therefore, null hypothesis for item2 can be accepted, there is no significant difference (very week) between high ability and low ability group. This item significantly does not discriminate well and it is very week for evaluation of the different levels, it should be eliminated from the examination.

When we compare items with respect to their discrimination indices, item 1 is very poor at discriminating (see table4); although 45.8% of those in the upper "high ability" group passed the item, almost same (46.5%) in the lower group passed the item. In tests of achievement or ability, negative D value would indicate a poor item in that those who scored most highly on the test overall were not likely to pass the item, whereas those with low overall scores were likely to pass the item (Kline, 2005). In the SOE, item 1 represents such discrepancy with respect to D discrimination parameter, its discrimination parameter is -0.01. This item does not differentiate high ability group and low ability group well with respect to D parameter. On the other hand it is poor but acceptable item according to its point biserial correlation coefficient rpb=0.52 and t-statistic=13.8. Remainder items show that, those who had the high exam scores were more likely to pass the items than examinees with low scores in whole examination. This situation indicates that, mentioned items have reasonable discrimination indices. The CTT statistics depicted that item 9 and item 11 have the highest rpbs, and D values (as parallel largest t-statistic too), these items best measure mathematics ability of students and best discriminate them. The mathematics test has a high alpha level which is determined as 0.844. In that case, there should be a low standard error of measurement (SEM) with respect to high internal consistency.



	p_high	p_low	Discrimination
item1	0.458	0.465	-0.01
item2	0.13	0.03	0.11
item3	0.67	0.08	0.59
item4	0.97	0.23	0.75
item5	0.94	0.23	0.71
item6	0.83	0.18	0.65
item7	0.76	0.09	0.67
item8	0.98	0.32	0.66
item9	0.92	0.11	0.81
item10	0.98	0.30	0.68
item11	0.96	0.18	0.77
item12	0.65	0.11	0.54
item13	0.79	0.08	0.70
item14	0.99	0.46	0.53
item15	0.77	0.16	0.61
item16	0.94	0.54	0.41
item17	0.88	0.16	0.72

Table 4:
probability of answering item correct for high ability group and low ability group and D-discrimination Indices.

Item 2 and item 16 have the lowest rpbs. If item they are deleted from the examination, minimum changes occur on Cronbach alpha. This situation indicates that effects of both questions are less than other items in overall mathematics examination to increase reliability.

In the comparison of difficulty indices, the following results are underlined; the closer the p value is

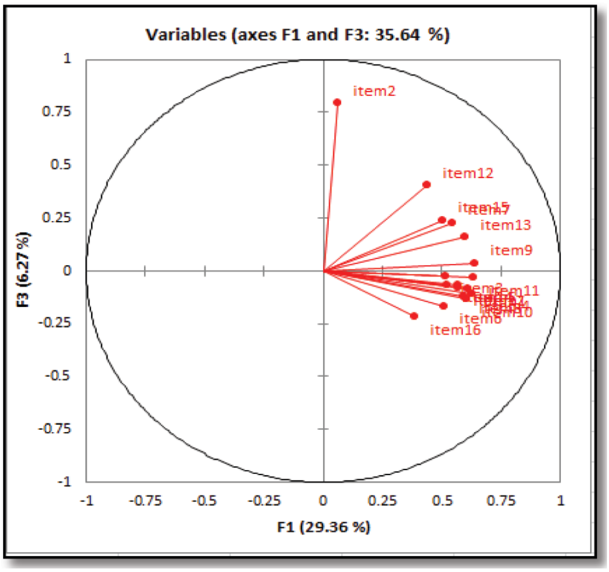


Figure 2 *First and second principal components (factors) and items distribution based on them.*

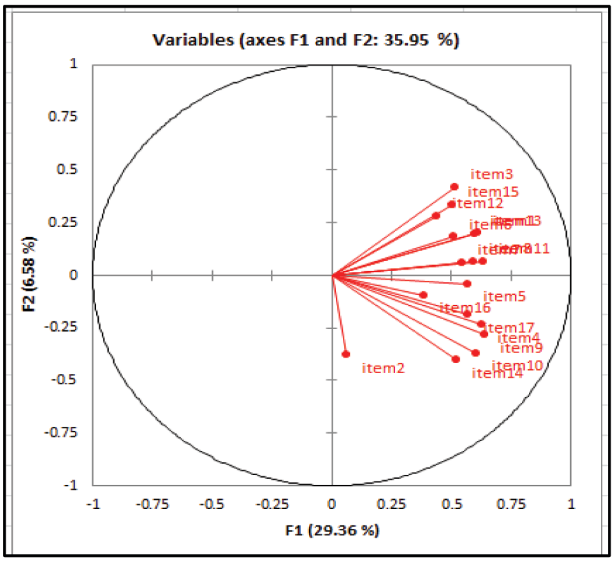


Figure 3 *Items distribution between first and third principal components.*

to 0.50, the more useful the item is at differentiating among test takers (Kline, 2005). Therefore, item2 and item 3 are the most difficult items (p2=0.11 and p3=0.28), while item 14 and 16 are the easiest ones (p14=0.78 and p16=0.77, almost same) with respect to their "p" difficulty indices (see table 3). Easiness of the item 14 and item 16 is very clear since 78% and 77% of the students are able to answer each of them correctly.

"Item 2" has the lowest corrected item total correlation and according to the PCA, only correlation between item 2 and third principle component "F3" is high (0.792), while this correlation is low with other components. Correlation between item1 and first principal component is very high. These cases are illustrated in figure 2 and figure 3.

Reliability Detection and Confidence Interval

All test scores are subject to measurement error. Because of measurement error, scores from alternate forms of a test, from one set of items to another, from one occasion to another, or, in cases where scores depend on human raters, from one rater to another, will not be perfectly consistent. In other words, scores on tests lack perfect precision due to the inevitable errors of measurement (Beckstead, 2013) .

In this part, standard error of measurement is calculated for mathematics examination. Therefore, confidence interval determined for each observed score. Steps of evaluation are expressed in the following way.

Starting point is using reliability coefficient and observed variance to determine SEM. Confidence intervals vary with reliability and observed variance is simply the variance of the observed test scores, when reliability coefficient is known, it is possible to estimate error variance:

Error Variance= (Observed Variance)*(1-Reliability Coefficient), and it can be denoted as:

$$\sigma_e^2 = \sigma_x^2 \left(1 - \rho_{xt}^2\right) \tag{11}$$

As the square root of a variance is a standard deviation, the square root of an error variance is a standard error of measurement (SEM). As we can use any standard error to build confidence intervals, we can use the SEM to build confidence intervals around an observed score (Suen, Principles of test theories, 1990). Given a fixed value of sample standard deviation of test scores, the higher the reliability of the test, the smaller the SEM (Ji Zeng,Adam Wyse, 2009). In this test SEM is calculated as SEM=1.6836. The confidence range is symmetric about the estimated true score. Consequently, if x is the observed score; the true score of a student lies between "x+1.6836 " and "x-1.6836", in the mathematics test. More precisely confidence interval can be determined as, "95% of probability the true score lies between "x+2*1.6836 " and"x-2*1.6836 " ".

Conclusion

This paper examined the behavior of item and person statistics empirically obtained from the CTT measurement frameworks. Before detecting internal consistency, unidimensionality of the mathematics examination is presented. Therefore, it is proved that, mathematics examination measure only one trait of students. In the introduction part, research questions were defined clearly. Overall, the findings from this investigation are presented here with respect to order of given questions.

CTT item discrimination indices are evaluated with point biserial correlation (rpb) and D-discrimination index. Except item 1 and item 2, remainder items have moderately high discrimination indices. This result indicates that, item1 and item 2 discriminate very poor but other 15 items discriminate well high ability groups and low ability groups.

CTT item difficulty indexes (p) are determined. Item 2 and item 3 are most difficult items, while item 14 and item 16 are the easiest ones.

Correlation between rpb and D-discrimination is 0.60. Both discrimination indices give parallel results. They discriminate items almost same, but interpretations are more meaningful with point biserial correlation because of support of t-test.

Reliability “internal consistency” is calculated using Cronbach alpha formula. Alpha is 0.844 which shows high internal consistency. When Item 2 and item 16 are deleted, minimum changes occur on alpha but deleting the items increases the alpha generally.

SEM is found as 1.6836. Confidence interval identified as:

$$x - 2*1.6836 \leq \text{true score} \leq x + 2*1.6836$$

where x is observed score. This study is mainly concentrated on CTT assumptions. Future study aims using more sophisticated and developed Item Response Theory methods to analyze instructions.

References

Beckstead, J. W. (2013, July). On Measurements and their Quality: Reliability-History, Issues and Procedures. International Jounal of Nursing Studies, 50(7), pp. 968-973.

Brown, J. (2001). Statistics Corner: Questions and Answers About Language Testing Statistics: Point--Biserial Correlation Coefficients. Shiken: JLT Testing & Evlution SIG Newsletter, 5(3), 13-17.

Cohen, B. H. (2008). Explaining Psychological Statistics. Hoboken, NewJersey, USA: John Wiley & Sons.

Cortina, J. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. Journal of Applied Psychology, 78(1), 98-104.

Cronbach, L. (1951). Coefficient Alpha and the Internal Structure of Tests. Psychometrica, 16(3), pp. 297-334.

Cronbach, L.J.; Shavelson, R.J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement, 64(3), 391-418.

Erguven, M. (2012). Comparison of the Efficiency of Principal Component Analysis and Multiple Linear Regression to Determine Students’ Academic Achievement. IEEE AICT , pp. 1-5.

Field, A. (2005). Reliability Analysis. In Discovering statistics using SPSS. London.

Gardner, P. (1996). The dimensionality of attitude scales: a widely misunderstood idea. International Journal of Science Education, 18(8), 913-919.

Ji Zeng,Adam Wyse. (2009, September 25). Introduction to Classical Test Theory. Michigan, Washington, US.

Jolliffe, I. T. (2002). Principal Component Analysis,. USA: Springer-Verlag New York Inc.

Kline, T. J. (2005). Classical Test Theory Assumptions, Equations, Limitations, and Item Analyses. In T. J. Kline, Psychological Testing (pp. 91-106). Calgary, Canada: SAGE Publications.

Linn, R. (2010). Educational Measurement: Overview. In P. Peterson, E. Baker, & B. McGaw, International Encyclopedia of Education (pp. 45-49). Elsevier.

Liu, F. (2008). Comparison Of Several Popular Discrimination Indices Based On Different Criteria And Their Application In Item Analysis. Athens, Georgia, US.

Mango, C. (2009). Demonstrating the Difference between Classical Test Theory and Item Response Theory Using Derived Test Data. The International Journal of Educational and Psychological Assessment, 1(1), 1-11.

N.M. Webb, R. S. (2006). Reliability Coefficients and Generalizability Theory (Elsevier B.V ed., Vol. 26). USA.

Privitera, G. J. (2011). Statistics for the Behavioral Sciences. USA: Sage Publications Inc.

Ronald K. Hambleton, R. W. (n.d.). An NCME Instructional Modulo on Comparison of classical test theory and item responce theory and their applications to test development. University of Massachusetts.

Sharkness, J.; De Angelo, L. (2010). Measuring Student Involvement:A Comparison of Classical Test Theory and Item response theory in the Cosnstruction of Scales from Student Surveys. Higher Education Research Institute, 52, 480-507.

Steyer, R. (1999). Steyer, R. 1999, Classical (Psychometric) Test Theory. Jena, Germany. Jena, Germany. Suen, H. K. (1990). Principles of test theories. New Jersey: Lawrence Erlbaum Associates ,Inc., .

Varma, S. (2013). Preliminary Item statistics Using Point Biserial Correlation and p-Values. Retrieved October 2013, from Educational Data Systems: http://www.eddata.com/resources/publications/EDS_Point_Biserial.pdf

Webb, N.M.; Shavelson, R.J.;Haertel, E.H. (2006). Reliability Coefficients and Generalizability Theory (Elsevier B.V ed., Vol. 26). USA.