Using Concepts of Control Charts to Establish the Index of Evaluation for Test Quality

CHIU-YAO TING[1]
Graduate School of Management
National Yunlin University of Science and Technology
123, Section 3, University Road, Douliou, Yunlin 640, Taiwan, R.O.C.
TAIWAN
CHWEN-TZENG SU[2]
Department of Industrial Management
National Yunlin University of Science and Technology
123, Section 3, University Road, Douliou, Yunlin 640, Taiwan, R.O.C.
TAIWAN
E-mail: g9120802@yuntech.edu.tw

Abstract: The purpose of this research employs concepts of control charts to establish a new evaluation standard for a test quality index. First, this study establishes criteria for assessing test quality by exploring the item discrimination index (D) in the general test areas. Further, data of this one index are normalized, and standard deviation is applied for the control limit of the control chart. Thus, the zone rules for the standardized D (SD) control chart proposed in this study are found to be an excellent evaluation method for test quality from the perspective of control charts.

Key-Words: Test quality, control charts, item discrimination, standard deviation, control limit, zone rules

1 Introduction

A test is a means to appraise a person's ability or knowledge in a certain field [1]; whereas, item analysis is a process of examining the response of an examinee toward each test question to determine test quality [2]. Since test questions can be used repetitively, the overall test quality will be improved and can serve as an item bank by modifying part of the examination questions for each application. Excellent test questions will be reserved and examination questions of poor quality will be corrected or deleted through test quality analysis. In this way, good test questions will be collected for future use.

Academic studies on the methods for evaluating the quality of each examination question of the Four-Year of Institute of Technology, and for the Two-Year Junior College Examination in Taiwan have been conducted. Formulation of questions for each subject in the entrance examination is based on knowledge competence in a two-way specification table. An examination paper can be comprised of several types of questions and the examination paper to be discussed here for Production Management constitutes units of Supply Chain Management (SCM), and Quality Management etc. Thus far, there have been few

studies related to unit-based test quality assessment and analyses.

A major goal of statistical process control (SPC) is to quickly detect the occurrence of process shifts so examination of the process and corrective action may be undertaken before many conforming units are manufactured. The control chart is an on-line process-monitoring technique widely used for this object. The control chart may also be used for parameters of a production process, and, through this information, determine process capability. Control charts may also provide information useful in improving the process [3]. Under the test quality application scenario, if the test quality is in control, as long as the D of items for the examination paper plots within the control limits, the test quality is assumed to be in control. However, a point that plots outside the test quality is out of control, and investigation and corrective action are required to find and remove the assignable cause of this behavior. In this study, the assignable cause variation means the examination paper could not discriminate among the ability of testees, for example, the difficulty of the examination paper is either high or low. This assignable cause variation may influence the testees' entrance into the school, and further, it may influence the professional image of the test organization.

Item discrimination index, as analyzed in this study, is one of the quality characteristics of a variable. The quality of the test questions is evaluated by this one test quality index in this article. As well as exploring the test quality indices in accordance with the evaluation standards of general tests, data of this one test is normalized and standard deviation is used for the control limit of the control chart for establishing an assessment criterion for test quality, we called the assessment criterion is the standardized D (SD) control chart.

This following section includes item discrimination and the concepts of control charts, methods, numerical results, improved test quality based on results and conclusions.

2 Item Discrimination and Concepts of Control Charts

Item discrimination refers to how well an item can discriminate or distinguish among testees that differ on the construct being measured by the test. Probably the most popular method of calculating an index of item discrimination is based on the difference between two groups [4], for example, high ability testees and low ability testees. Item discrimination is computed for each group separately, and these are labeled P_T and P_B. Item discrimination is calculated by the following formula [5]:

$$D = P_T - P_B \quad (1)$$

where

D =discrimination index

P_T =proportion of testees in the top group getting item correct

P_B =proportion of testees in the bottom getting item correct

A typical control chart is shown in Figure 1, which is a graphical display of a quality characteristic measured or computed from a sample versus the sample number or time [3]. The chart consists of a central limit (CL), upper control limit (UCL) and lower control limit (LCL). The control chart is expressed by the general formulas from equations (2) through (4). Assume X is the specimen statistic of quality characteristic, μ_x is the mean of X, and σ_x is the standard deviation. The central line is mapped in the control chart. The central line refers to the process characteristic with no abnormal variables. K is the distance from the control line to the centerline, expressed by a certain multiple of the standard deviation.

$$UCL = \mu_X + K\sigma_X \quad (2)$$

$$CL = \mu_X \quad (3)$$

$$LCL = \mu_X - K\sigma_X \quad (4)$$

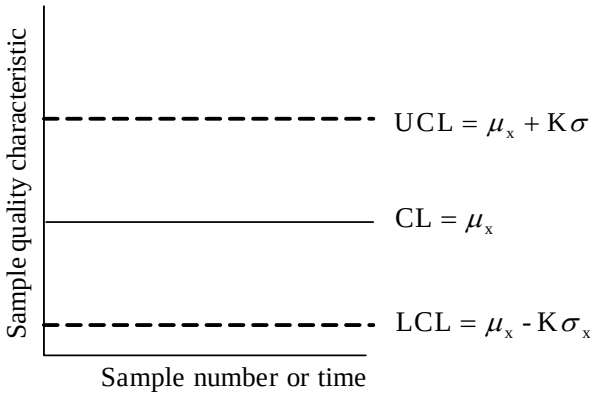


Fig. 1. A typical control chart

There are two types of control charts: variables control chart and attributes control chart. If the quality characteristics can be quantified and continuous, they belong to variables. Many quality

characteristics can be expressed by variables. This type of control chart is called variables control chart [3]. In this study, the D belongs to a variable.

Control limits are confined to the natural variations of a process. After considering the process standard deviation and natural tolerance limits of the process, $\pm 3\sigma$ (standard deviation) is usually used as the upper (UNTL) and lower natural tolerance limits (LNTL); Specification limits are specified by the management, customers or even the product developers. As a whole, there is no numerical or statistical correlation between the control limits and specification limits. Only when individual observations are mapped, can specification limits be significant. The relationships among control limits, specification limits and natural tolerance limits are shown in Fig. 1 [3].

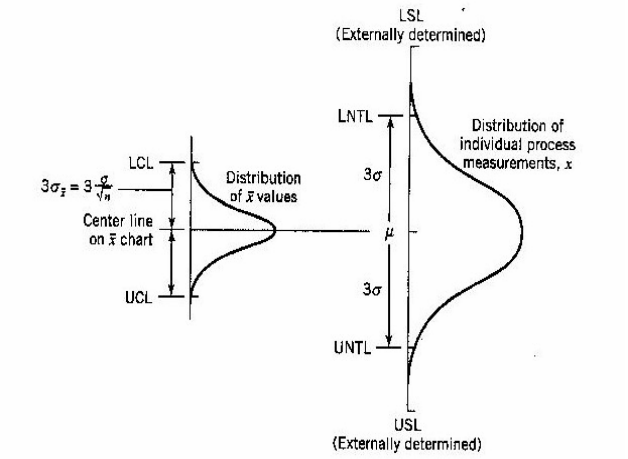


Fig. 1 Control limits, specification limits, and natural tolerance limits

Two groups of control limits may be applied for the control chart in this study. The external limit UNTL and LNTL are 3σ, the most common action limit. The internal limit, which is a certain multiple of the standard deviation, is decided according to the purpose of application and is termed the warning limit. Using the warning limit may heighten the sensitivity of the control chart because parametric deviation during the process can be observed rapidly [3]. Besides, the sample size in this study is n=1, that is, the sample consists of an individual observation, so we use the individual control chart evaluate the test quality.

3 Methods

The main purpose of this study is to analyze the test quality of each unit in Production Management. The statistical analysis system, Minitab14.0 and Eviews 5.0 are adopted.

3.1 Test of Normalization, Randomness, and Independence for Data

When a control chart is applied to evaluate test quality, the data under study must be subject to normal distribution, randomness and independence and be tested so quality control can be conducted by the characteristics of the control chart.

For the normalized test, we can use the normality test (Kolmogorov-Smirnor goodness-of-fit test) using software (for Minitab) to understand whether the data (seven units for items) follows normal distribution. The hypothesis test of data with $\alpha = 0.05$ is as follows:

H_0: True cdf is the normal distribution with estimated parameters.

H_1: True cdf is not the normal distribution with estimated parameters.

For the randomness test, we can use runs tests using software (for Minitab) to see whether the data residual follows randomness. The hypothesis test of data with $\alpha = 0.05$ is as follows:

H_0: Data is random.

H_1: Data is not the random.

For the independence test, we used the autocorrelation function (ACF) with time series using software (for Eviews) to test the independence of data. The hypothesis test of data with $\alpha = 0.05$ is as follows:

H_0: Data is independent.

H_1: Data is not the independent.

3.2 Establishing the Standard of Evaluation for Test Quality

In this study, quality evaluations divide into two conditions based on the item discrimination index. The assessment criteria commonly used in the test field are adopted with condition D presumed, resulting in the individuals control chart. The second condition is the evaluation standards presented in this study; the standardized D (SD) control chart, also termed as condition Z_D, is utilized for judging test quality.

For condition D, the literature survey is conducted and the D established by previous scholars. It is used as the norm to determine the quality of a test as well as the basis of mapping a control chart. The evaluation criterion mostly used for general tests is

utilized in the first circumstance and the judgment standard is described as follows [6]: D<0.2: Poor; 0.2<D<0.3: Acceptable; 0.3<D<0.4: Good; 0.4≦D: Excellent. Based on the index values above the upper bound (UB) of individuals control chart being 0.4, the middle bound (MB) is assumed to be 0.3, and the lower bound (LB) is assumed to be 0.2 in this study. Items above the UB are excellent. Items below the LB are poor. Items between the UB and the MB are of good quality. Items between the MB and the LB are of acceptable quality.

As for condition Z_D, the standard deviation in the control chart is applied as the evaluation criteria, resulting in an SD control chart. Therefore, D has to be converted to standard normal Z for mapping of the control chart.

The main purpose of this study is to construct a new evaluation standard for test quality indexes. The reason for using a control limit of 3σ is 99.73% of items are within the area of the mean $\pm 3\sigma$. Therefore, six quality areas are classified as follows, and we term them zone rules for control charts of test quality in this study: zone A, including PA ($+1\sigma$) and NA (-1σ): Excellent, zone B, including PB ($+2\sigma$) and NB (-2σ): Acceptable and zone C, including PC ($+3\sigma$) and NC (-3σ): Good.

Here 1σ serves as the warning limit because 68.28% of the values are within the area of 1σ. A comparison of the Z scores implies a discrimination range from 0.1587 to 0.8413 at 1σ and the discrimination of most of the items are covered in this area. If 2σ is used, discrimination will range from 0.0228 to 0.9772. Thus, 1σ serves as the warning limit in this study.

4. Numerical Results

This section will be explained by describing basic information, data processing and control chart of test quality.

4.1 Description of Basic Information

Examination paper preparers need to consider the contents of teaching materials and principles of teaching and use these two factors as the scheme for a two-way specification table. Classification of teaching objectives is usually based on the six cognitive levels proposed by Bloom et al. (1956) [7]. A two-way specification table properly facilitates content and cognitive distribution in an examination paper, so the arrangement of items becomes better structured through a two-way specification table.

A two-way specification table is used as the structural design for a test. It describes the necessary contents and ability to be measured in a test, and is used for reference in formulating questions for a test. It is based on the two axes of teaching objectives and learning contents, explaining various measurement goals respectively. Establishing a two-way specification table may help to formulate questions for tests, clarifying the relationship between teaching objectives and learning contents to ensure the contents of teaching materials can be reflected by the tests and the expected learning results can be measured. 4,544 students taking the test on Production Management are the objects of study in this article and belong to a large sample. In addition, discrimination D is between 0~1, if D follows normalcy, they can obtain standard normal Z scores for evaluating test quality. Every item is further classified by seven units in the two-way specification table of Production Management shown in Table 1.

Table 1. Two-way specification table of Production Management

Unit \ Cognitive	knowledge	comprehension	application	synthesis and analysis
1 Introduction		1,2,3,4,5,6,7,8	10	9
2 Forecasting		11,12,14,15,16, 17,19,20		
3 Capacity Planning		13,18		
4 SCM	22	23	21,24,25	
5 Scheduling		26,27,28,29,30,31,33		
6 Inventory Management		32,36	34,35,37,38	39,40
7 Quality Management				41,42,43,44,45,46,47,48, 49,50

The discrimination (D) of seven units for all 50 items, item size, means, and standard deviations are rearranged in Table 2.

Table 2. D of seven units for items

Introduction		Forecasting		SCM		Capacity Planning		Scheduling		Inventory Management		Quality Management	
Item Number, i	D_i	Item Number, i	D_i	Item Number, i	D_i	Item Number, i	D_i	Item Number, i	D_i	Item Number, i	D_i	Item Number, i	D_i
1	0.39	11	0.38	13	0.16	21	0.45	26	0.30	32	0.50	41	0.29
2	0.14	12	0.27	18	0.10	22	0.65	27	0.11	34	0.13	42	0.43
3	0.38	14	0.20			23	0.61	28	0.31	35	0.56	43	0.40
4	0.31	15	0.04			24	0.48	29	0.31	36	0.25	44	0.46
5	0.25	16	0.18			25	0.61	30	0.32	37	0.32	45	0.39
6	0.23	17	0.27					31	0.26	38	0.28	46	0.23
7	0.17	19	0.09					33	0.39	39	0.31	47	0.10
8	0.20	20	0.31							40	0.33	48	0.01
9	0.12											49	0.30
10	0.31											50	0.07
Item Size (N)	10		8		2		5		7		8		10
Mean ($\overline{D}_i$)	0.25		0.2175		0.13		0.56		0.2857		0.335		0.268
StDev (σ_{D_i})	0.0955		0.1134		0.0424		0.0889		0.0866		0.1368		0.1607

4.2 Data Processing

The standardized discrimination (SD) of seven units for items are summarized in Table 3. Z_D means to convert discrimination into a standard normal Z.

Table 3. Calculations for the SD of each unit for items

Introduction		Forecasting		SCM		Capacity Planning		Scheduling		Inventory Management		Quality Management	
Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}	Item Number, i	Z_{D_i}
1	1.4667	11	1.433	13	0.707	21	-1.2376	26	0.1652	32	1.2061	41	0.1369
2	-1.1524	12	0.463	18	-0.707	22	1.0126	27	-2.0293	34	-1.4985	42	1.0081
3	1.362	14	-0.1543			23	0.5626	28	0.2807	35	1.6447	43	0.8214
4	0.6286	15	-1.5653			24	-0.9001	29	0.2807	36	-0.6213	44	1.1948
5	0	16	-0.3307			25	0.5626	30	0.3962	37	-0.1096	45	0.7592
6	-0.2095	17	0.463					31	-0.2968	38	-0.402	46	-0.2365
7	-0.8381	19	-1.1243					33	1.2047	39	-0.1827	47	-1.0454
8	-0.5238	20	0.8157							40	-0.0365	48	-1.6055
9	-1.362											49	0.1991
10	0.6286											50	-1.2321

Note: $Z_{D_i} = \dfrac{D_i - \overline{D_i}}{\sigma_{D_i}}$

4.3 Test of Normalization, Randomness, and Independence for Seven Units

We suppose that all students are independent in this study. Results of the normalization, randomness, and independence for seven units are presented in Table 4. As seen in Table 4, the normalize test of seven units for items point out no reject H_0 (p-value> $\alpha = 0.05$), thus, we rationally suppose seven units for items follow the normal distribution; with the randomness test of seven units for items indicating no reject H_0 (p-value> $\alpha = 0.05$), which means we could rationally suppose the data follows randomness; with ACF of seven units for items, this points out the observed data has no serious autocorrelation (p-value> $\alpha = 0.05$), which means the data follows independence. To sum up, the data follows normalization, randomness, and independence. In other words, the D of the data follows the assumption of the control chart. Thus, the data in this study could use the control as the instrument for test quality analysis.

Table 4. Test of normalization, randomness, and independence for seven units

Method	Unit	Observations	p-value
Normality test (K-S)	Introduction	10	>0.15
	Forecasting	8	>0.15
	SCM	2	>0.15
	Capacity Planning	5	>0.15
	Scheduling	7	0.094
	Inventory Management	8	0.094
	Quality Management	10	0.15
Runs test	Introduction	10	0.574
	Forecasting	8	1.000
	SCM	2	[]
	Capacity Planning	5	0.513
	Scheduling	7	0.224
	Inventory Management	8	1.000
	Quality Management	10	0.206
ACF	Introduction	10	0.228
	Forecasting	8	0.864
	SCM	2	[]
	Capacity Planning	5	0.117
	Scheduling	7	0.256
	Inventory Management	8	0.171
	Quality Management	10	0.085

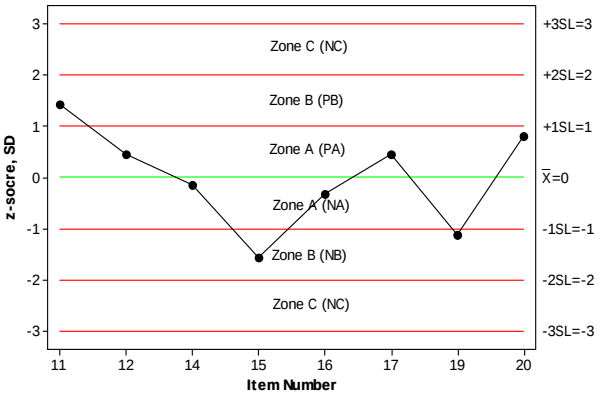
Note: because the observations of SCM unit are too small (two items) so we cannot obtain the p-value.

To sum up, the data follows normalization, randomness, and independence. In other words, the D of the data follows the assumption of the control chart. Thus, the data in this study could use the control as the instrument for test quality analysis.

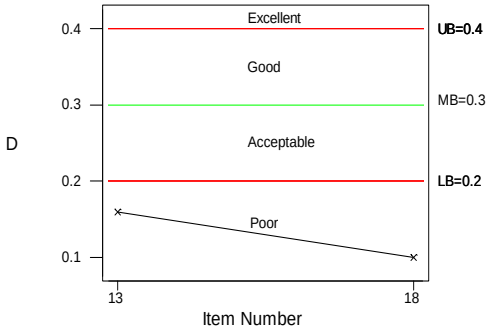
4.4 D and SD Control Charts of Test Quality for Units

In designing a control chart, we must specify both the sample size to use and control limit. In this study, the sample size is 1, total numbers of items is 50, and the control limits are ±3.The x-axis represents the item number, and the y-axis represents the D-index of each item, with the D control chart and SD control chart of the seven units using the Minitab individuals control chart. There are seven units of data for analysis in this study. Take the Forecasting unit, Capacity Planning unit, Inventory Management unit, SCM unit and the Scheduling unit as examples with the D control chart and the SD control chart being mapped from Fig. 2a to Fig. 2j.

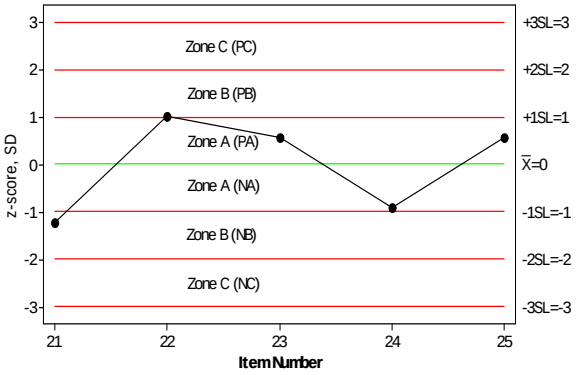
(a) D control chart of Forecasting unit



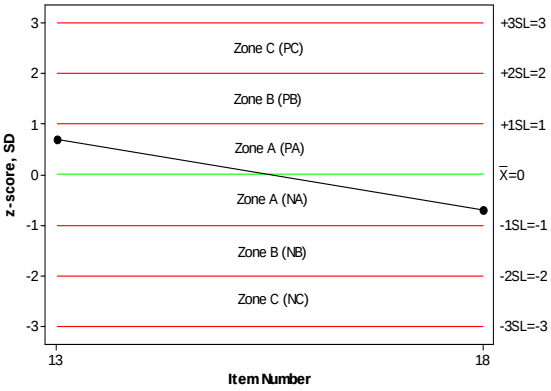
(b) SD control chart of Forecasting unit



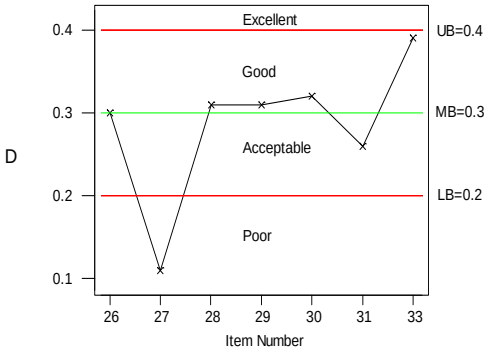
(c) D control chart of SCM unit



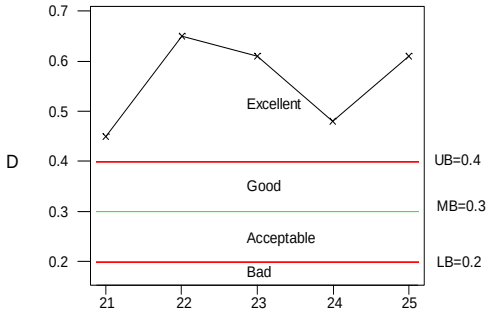
(f) SD control chart of Capacity Planning unit



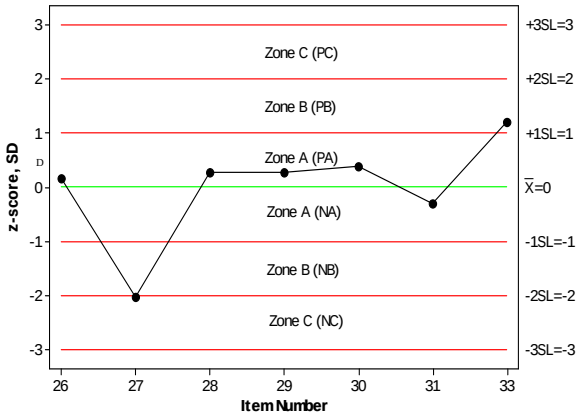
(d) SD control chart of SCM unit



(g) D control chart of Scheduling unit

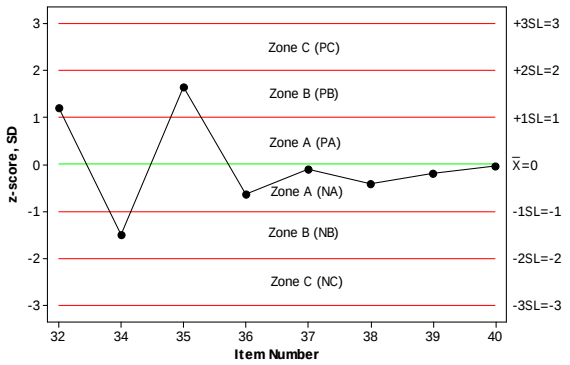


(e)D control chart of Capacity Planning unit



(h) SD control chart of Scheduling unit

(i) D control chart of Inventory Management unit



(j) SD control chart of Inventory Management unit

Fig. 2. D and SD control chart of each unit

The results of the D and SD control charts are analyzed and summarized as follows:

For the D control chart, which is another index for quality assessment in the current testing field, there are 13 test questions in the examination paper having "Poor" quality, including item numbers 2, 7, 9, 13, 15, 16, 18, 19, 27, 34, 47, 48 and 50. The other test items are within the "Acceptable" areas, including the "Acceptable," "Good" and "Excellent" quality areas, implying the overall test is at least acceptable.

As for the SD control chart (Figure. 2d), there is one item in the NC area, which is item number 27. For the seven units, the test quality of six units is within zone A or zone B, and except for the Scheduling unit, the six units are within the "Acceptable" areas.

According to the above and Figure 2d, the Scheduling unit includes the zone A (Excellent), zone B (Acceptable) and NC area (Good), located within $\pm 3\sigma$, revealing the discrimination is lower than in zone B, but does not exceed 3σ, rendering an acceptably good quality.

5 Improvement in Test Quality based on Results

The improvement in test quality based on study results is explained by the D and Z_D in the general test and the SD control chart in this study. For condition D, "Acceptable" discrimination of most of the units implies a fair test quality, which is not the same as under the D condition possibly causing a misjudgment of test quality.

As for the condition of Z_D, the quality of all units are within the A area with no unit greater than 3σ, which indicates the quality of item data analyzed in this study is excellent, with no poor test quality for any unit.

This finding of this study that the D condition, there are 24 items with D values below 0.30 (see Table 2). It implies that nearly half of all items should be carefully reviewed and possibly revised or deleted [4], as they may result in increased risk of false alarms, perhaps requiring future research exploration. In contract, for the Z_D condition, all seven units are evaluated as at least having acceptably good quality. Thus, the SD control chart with 3σ as the control limits and 1σ as the warning limits not only has better quality evaluation capability than the D condition, but also displays the test quality of each unit in a clearer and more specific way. Therefore, the SD control chart can serve as a highly effective approach to test measurements.

6 Conclusions

One important characteristic of item analysis in current test theory is to present the test questions by charts and graphs for easy understanding. The purpose of this article is to examine the test quality of a unit-base subject by using the quality control features of a control chart. The test data to be analyzed in this study is the level of discrimination, which is one of the quality characteristics of a variable. As a result, the overall ability is examined, and the quality control features of a control chart are applied with related statistical methods being integrated for measuring test quality in the field of Production Management.

Diagrammatic control charts and statistic-based quality management characteristics are applied for evaluating unit test quality. This study not only provides a way to test quality analysis, but also helps persons who formulate questions for tests to be aware of key points when preparing for examination papers. Suggestions such as ability-oriented items are offered for improving test quality.

As well as complying with evaluation standards for general tests to explore the measurement standard for the test quality index, a standard deviation is also used as the control limits of the control chart, serving as a criterion for evaluating test quality. This study's results show the evaluation standard adopted in this article cannot only help to understand the quality level more clearly, but also show the status of test quality more specifically. Thus, the standard deviation control approach proposed in this study is found to be an excellent evaluation method for test quality from the control chart. A consistent evaluation criterion for test quality can be obtained by the SD control chart.

For entrance examinations, questions on each subject should be formulated in a way similar to the principle mentioned in this article. Therefore, the results of this study can be applied to test quality analysis of various subjects, including Production Management or entrance aptitude tests.

References:

1. Bloom, B., M., Englehart, E., Furst, W. Hill and D. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals,* Handbook I: Cognitive Domain, White Plains, NY: Longman, 1956.
2. Brown, H. D., *Principles of Language Learning and Teaching* (3rd ed.), Englewood Cliffs, NJ: Prentice Hall Regents, 1994.
3. Hopkins, K. D., *Educational and Psychological Measurement and Evaluation* (8th ed.), Boston: Allyn & Bacon, 1998.
4. Johnson, A. P., Notes on a suggested index of item validity: The U-L index, *Journal of Educational Measurement*, This is a seminal article in the history of item analysis Vol.42, 1951, pp. 499-504.
5. Mehrens, W. A. and I. J. Lehmann, *Measurement and Evaluation in Education and Psychology* (3rd ed). New York: Holt, Rinehart, and Winston, 1984.
6. Montgomery, D. C., *Statistical Quality Control* (5th ed.), John Wiley and Sons, 2005.
7. Reynolds, C. R., R. B. Livingston and V. Willson, *Measurement and Assessment in Education*, Boston: Pearson Education, 2006.