

Research Paper

Beyond Measurement: An Evaluation Framework for  
Employee Performance Measurement Systems

**Ugur Cem YILDIZ**

Organizational Development

TOFAS Human Resources Directorate

Address: Tofas Fabrika, Yalova Yolu, 10.km, 16369, Bursa, Turkey

Tel: 90 (224) 261 03 50

Fax: 90 (224) 257 05 80

## Research Paper

# Beyond Measurement: An Evaluation Framework for Employee Performance Measurement Systems

### ABSTRACT

**Purpose:** Although quite a lot research effort has been devoted to establishing more efficient and effective performance management systems for measuring the employees' performance better, the issue of evaluating performance measurement results in a systematic manner remained relatively untouched. This paper argues problems of evaluating results of performance measurement system of Tofas Turk Otomobil Fabrikasi and discusses possible ways of handling the evaluation process methodically. **Methodology:** The study is based on empirical investigation of Tofas' experience and the research done on evaluating educational performance. **Findings:** The paper presents empirical evidences of Tofas' case and identifies relevance of educational performance evaluation issue to the discussion. By combining both, it proposes a set of statistical approaches for a more systematic evaluation process. **Originality / value:** The findings may especially prove valuable for HR specialists who practice objective-based performance appraisal.

**Keywords:** Performance Management, Performance Measurement, Performance Evaluation, Relative Evaluation

## INDEX

<i>Introduction</i>	4
<b>Performance Management</b>	4
<b>Performance Measurement System of Tofas</b>	5
<b>Factors of Relativity</b>	8
<b>The Relative Evaluation Principle</b>	12
<i>Educational Performance</i>	12
<b>Evaluation Frameworks for Educational Performance</b>	12
<b>An Example Methodology (Istanbul University)</b>	13
<b>An Example Methodology (OSYM)</b>	14
<i>The Proposed Methodology</i>	17
<b>Introduction</b>	17
<b>Step 1: Grouping</b>	17
<b>Step 2: Clustering</b>	18
<b>Step 3: Transforming</b>	19
<b>An Alternative Methodology</b>	24
<i>Conclusion</i>	25
<i>References</i>	26
<i>Indices</i>	26

## **Introduction**

### ***Performance Management***

Performance management is a relatively new concept to the field of management. As a distinct and formal management procedure used in the evaluation of work performance, it dates not more than 60 years ago. At first, it began as a simple method of income justification. Later, its developmental importance was understood, and therefore, along this short history, quite a lot effort has been devoted to establishing more efficient and effective performance management systems.

By definition, “Performance Management” concept includes two sub-concepts: “Performance Measurement”, which is the activity of observing the performance of employees and quantifying their achievements according to observation results, and “Performance Evaluation (Assessment)”, which is the act of deriving conclusions from the measurement results.

Literature bestows a quite solid background for the measurement sub-concept. (Neely, 2005) It is possible to find a lot of research on two typical measurement systems, which are competency-based and objective-based systems. Objective-based measurement systems have especially been popular since 50s after Peter Drucker’s book “The Practice of Management”, which introduced Management by Objectives (MBO) concept. Having been derived from the “MBO”, so many popular approaches evolved. One of the most popular of them is Kaplan & Norton’s “Balanced Scorecard” introduced as a management tool in 1991.

As for evaluation sub-concept, however, it is not possible to come across with as much work as in the case of measurement sub-concept and most relevant work reside in another part of the literature, which is educational performance evaluation. Today, many large companies employ relative evaluation methods together with hybrid (both objective and competency based) measurement systems. However, these methods generally rely on discretion of managers rather than a systematic approach. For example, a USA based technology company with 350,000 employees worldwide employs an objective-based performance management system and requires a relative evaluation. The evaluation process requires evaluating an employee in comparison with the others, therefore the method is called “Relative Evaluation”. HR asks managers to transform the measured scores of employees into relative results over a referential scale. Most large companies assume a similar approach

but some bring additional restrictions. For example, another USA based transnational company in personal products industry employs a system similar to the aforementioned company's but it requires that each department's evaluation must comply with a pre-determined distribution, that is, percentage of high and low appraisals must be within a predefined range.

All of these practices actually put the applied performance management system in the role of a decision support system and leave the final decision about employees' performance evaluation to managers with the privilege of discarding measurement results. This is particularly because most companies believe in that the evaluation process, by nature, can't be made more systematic, it is not a mechanical process and requires a human touch. Some also allege that providing such flexibility to managers is not meaningless, they must have this power since they are completely accountable for performance of the groups (divisions) they are managing.

If managers utilize the performance management system intelligently and their final appraisal decisions do not conflict with the performance measurement results and the evaluation process runs transparently, then there appears no big problem. But if not, then the performance management system is put in question by employees. Above all, if these results directly reflect to compensation and career decisions then giving such flexibility to managers become even more problematic.

Having anticipated this problematic nature of evaluation issue, Tofas [<sup>1</sup>] decided to develop a systematic approach in evaluating measurement results before leaving its completely discretion-based performance management system and passing to an objective-based system in 2001. This article aims to discuss the problems of relative evaluation principle and explain Tofas' methodology to deal with these problems. The rest of the paper is organized as follows. Beforehand, basics of the system of Tofas are explained. Later the problem of evaluation and its similarities with educational performance issue is discussed. After referring to some example methodologies of educational performance evaluation, the proposed methodology is explained. The final section concludes the paper.

### ***Performance Measurement System of Tofas***

---

<sup>1</sup> Tofas Turk Otomobil Fabrikasi A.S., <http://www.tofas.com.tr>

Tofas' performance measurement system has two variants. One variant is based on a modified version of the Balanced Scorecard philosophy and it requires that the supervisor and the employee agree to what the employee will attempt to achieve in the period ahead, and that the employee accept and buy into the objectives. Employees create a scorecard by formulating a set of indicators assumed to outline their success along a period of 1 year. In line with the departmental targets, each employee sets objectives for the selected indicators at the beginning of the year and the scorecards are frozen after the approval of immediate supervisor until the evaluation period at the end of the year. Unless a major change happens in the job content, employees are not allowed to change their scorecards (indicators and objective values).

The other variant is designed for less qualified jobs, for which it is either difficult to formulize indicators or infeasible to monitor the indicators' progress. Most clerical and secretarial jobs fall into this category. If this variant is selected for an employee, the employee creates his/her scorecard by choosing a predetermined set of performance criteria, covering more than 10 different dimensions of work performance. Each criterion has equal weight and success at each criterion is evaluated over a verbally defined scale of 5 levels.

When the evaluation period arrives, numerical indicators' scores are automatically calculated by the related software according to the realizations of indicators, while the scores for non-numeric indicators are entered to the system by employees themselves. The cumulative score calculated from business objectives constitutes 70% or higher of the final score. After supervisors approve the values entered by employees, they enter their discretion scores (weighing 30% or less), so that the overall performance score for each employee is calculated.

After the scores are calculated, they are evaluated according to a relative success scale of 5 levels in order to be used in other HR systems. Those relative success levels are labeled as A, B, C, D and E, which are qualified as "extra-ordinarily successful", "very successful", "adequately successful", "need to be improved" and "unsuccessful" respectively. However the company requires that percentages of these success levels must conform to a pre-defined distribution (See Figure 1).

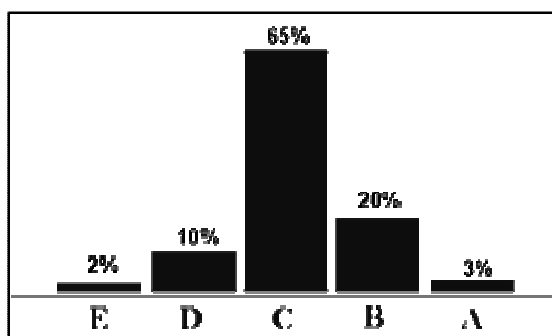


Figure 1 – The pre-defined distribution

The problem arises just at this point. Let’s clarify the issue with an example. Assume that score distributions of two departments in a company are as presented in Figure 2. Employees of Department A get 85 on average, while Department B’s employees get 125. X is the most successful employee of Department A while Y is the least successful employee of Department B. They both receive a score of 100. Does this mean that their performances are equal? Department B’s scores are distinctively higher. Does this mean that Department B is more successful than Department A and thus everyone in Department B must receive higher notes than anyone in Department A? Are the performance evaluations (scores) within these two departments comparable?

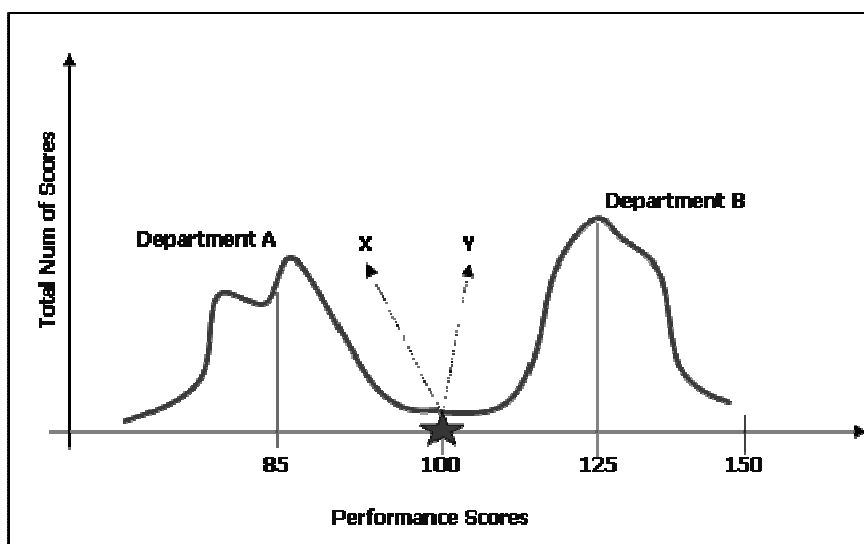


Figure 2 – Score distributions of two different departments

These are difficult questions, but our experience and analysis shows that the answer for all is “No!” Such radical variations generally do not have a legitimate explanation, and when we dive deep into the reasons, we see that different factors of variation (such as superintendents’ attitudes, departmental and job related differences) are at work. Due to these factors, the

measurement results of performance appraisal systems naturally become relative. This is like measuring performance of two athletes running on different tracks in different locations. If one is running on a sandy ground against the wind and the other is running on a regular track along the wind, it is necessary to consider these factors before making a final judgment about their performances.

### ***Factors of Relativity***

In Tofas' case, we identified two types of factors giving birth to relativity problem: Micro and Macro factors. (See Figure 3)

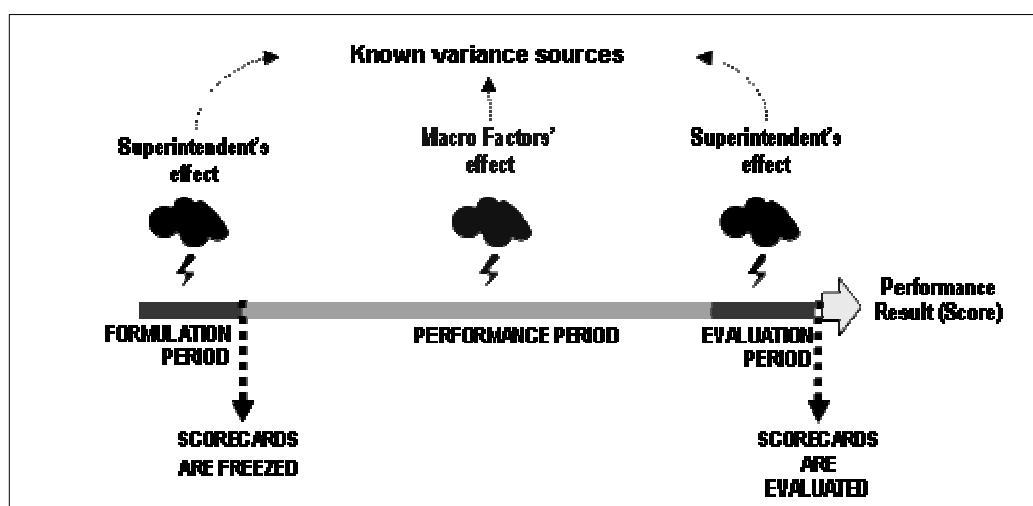


Figure 3 – Different effects are at work in different phases

### ***Micro Factors***

Micro factors originate from differences in natures of the jobs and from the way by which the superintendent and employee handle the system. According to our observations, among the micro factors, the most important to be considered is the superintendent's effect. As a necessity of the process, each employee prepares and evaluates his scorecard with his/her immediate superintendent. Therefore superintendents' attitudes and understanding of the system plays a significant role on both formation of scorecard and its evaluation. While one superintendent may be strict and more demanding, the other may be less demanding and more relax either in formulation or evaluation phases. Therefore, even if we ask two different superintendents to evaluate the same employee in a certain position, the measured performances will probably come out different due to those different personal attitudes.

The impact of personal attitudes can be observed well when 2005 scorecard results of 837 white collar employees in Tofas are examined. Although realization of targets in the company's scorecard was 89% in 2005, on average target realizations within the company was 107%. (See Figure 4) This means that the targets weren't properly deployed within the organization and employees and superintendents generally preferred to have more secure and less stretched targets, despite of a strictly managed deployment process.

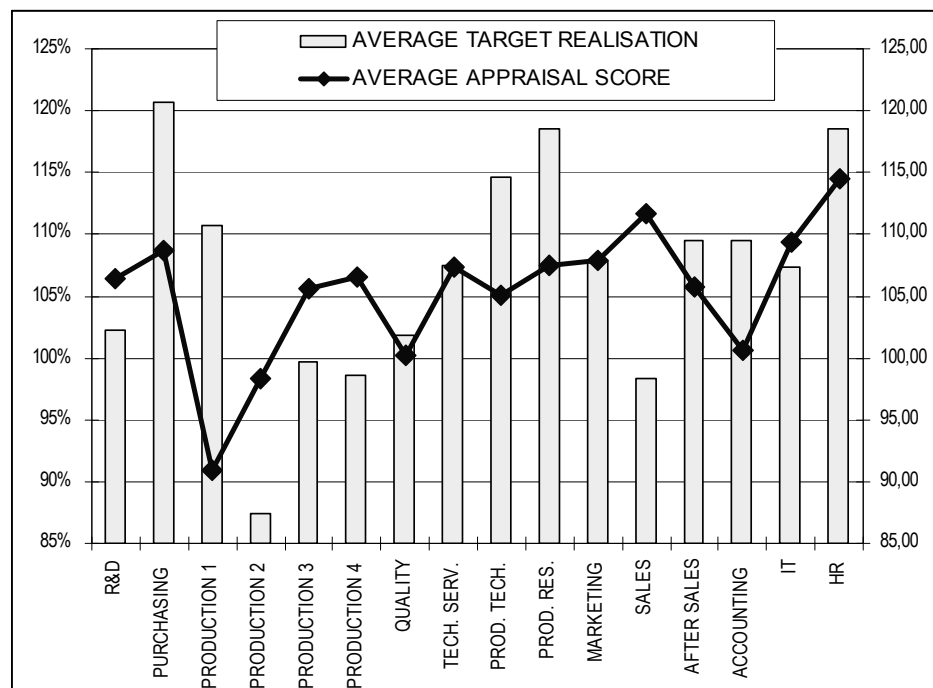


Figure 4 – Average target realization vs. average appraisal (discretion) score

Figure 4 provides another clue to us. In the departments where the target realizations on average were comparatively high the appraisal (discretion) scores on average were comparatively low and vice versa. This means that superintendents generally used appraisal (discretion) scores in order to normalize the target realization scores.

Another supportive argument resides in Figure 5. We observe little correlation (in general) between target realization scores and appraisal (discretion) scores, when we look at the distribution of scores within each department in detail. We can conclude from the differences in Figure 5 that most managers effectively use their appraisal (discretion) scores to manipulate the target realization scores.

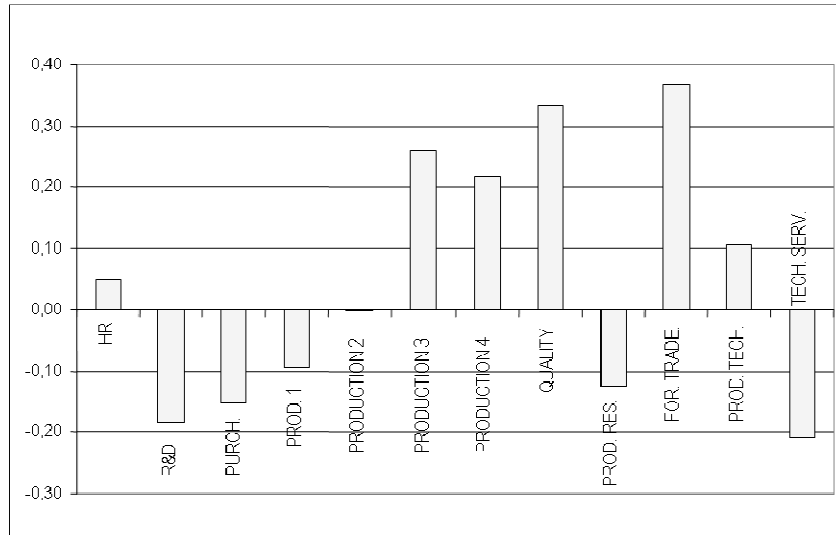


Figure 5 – Correlation between target realization scores and appraisal (discretion) scores

This conclusion is also supported by Figure 6, which presents the standard deviations of target realization scores and appraisal (discretion) scores. It can be seen that standard deviations of appraisal (discretion) scores were higher than those of target realization scores. This means that, in general, target realization scores were unable to differentiate employees very much or able to differentiate employees in a rather undesired way, hence the managers tried to balance it by assigning radically varying appraisal scores.

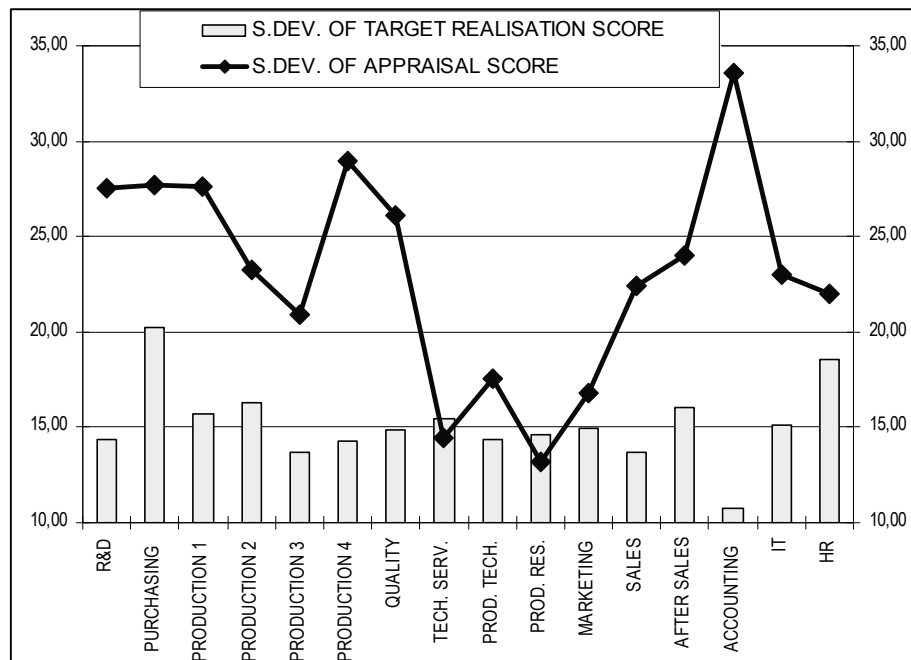


Figure 6 – Standard deviation of target realization scores and appraisal (discretion) scores

When Figure 5 and Figure 6 are interpreted together, it can be understood that in some departments (such as R&D, Purchasing,...etc) the appraisal (discretion) scores were used to pull the target realization scores down and in some other departments (such as Production 4, Quality...etc) the appraisal (discretion) scores worked in the opposite way.

Since the average appraisal score weights are quite high (on average 30%), we can conclude that the superintendents manipulated the final scores of employees in most departments considerably. Due to the significance of the superintendents' effect, we can conclude that performance measurement results in each organization box forms an independent population and hence these populations can not be compared directly.

### ***Macro Factors***

Other than the superintendents' effect, performance result of an employee is also affected from the macro (environmental) factors, which are out of control of that employee and his/her superintendent. (Figure 3) They affect all people in certain groups, functions or departments in the same manner. For example, a production halt due to a quality problem will be affecting almost everybody in Production Units (from production team leaders to procurement staff; from Production Unit 1 to Production Unit 4) in the same manner. Different departments may be subject to quite different macro factors, that is, a quality problem won't be affecting employees in Accounting Department as much as the ones in Production Units. Therefore, performance results of people in such different groups, functions or departments can not be compared directly. They must be treated as independent populations.

Figure 4 provides a good illustration of this case. While most support (staff) departments' target realization levels are quite high (such as Purchasing, Prod. Res., HR), most line departments' target realization levels are below average (such as Production departments) Although each departments' performance is very much related to each other, some specific factors affecting them created a significant difference in their overall performance results.

Although degree of effect of a specific macro factor may vary on personal basis within any specific group, function or department, we have to assume - for the sake of simplicity - that all people in a specific group are being affected equally from these macro factors. For example, an unexpected material procurement problem will certainly be affecting production

team leaders' scorecards directly and more profoundly while its effect will be relatively indirect and limited for technicians within the same production department.

### ***The Relative Evaluation Principle***

Since it is not possible to measure the performance of every employee in a corporation with the same standards and under the same circumstances, this relativity issue can't be avoided. Therefore most companies prefer relative evaluation methods in order to eliminate effects of relativity. Otherwise misuse and interdepartmental justice will become a serious problem to degenerate the performance management system in use since it is directly linked to compensation and other HR systems.

The problem of relative evaluation is that, eliminating effects of relativity in a systematic manner instead of relying on intuition of managers is a complicated process. It is required to convert reference free measurements into a comparable scale by taking many independent factors working differently for every employee into consideration. This problem of professional world is very similar to the problem of schools and universities, who would like to compare success levels of two students from two different classes or two different schools. Therefore we checked the research on evaluating educational performance.

## **Educational Performance**

### ***Evaluation Frameworks for Educational Performance***

The problem of relativity was first recognized in educational institutions. A note received by a student in a class depends on many factors such as the instructor's ability to motivate, ability to create a good learning environment, the methodology employed for measuring success levels of the students, the general success level of the class and the attitudes of the other students and the instructor. Since these factors affect measurement results detrimentally, measured success levels of two students in two different classes are theoretically incomparable.

In order to evaluate different measurements, many relative evaluation methods have been employed. (Ebel, 1965) Some of the most commonly used ones are:

1. **Evaluation according to the normal distribution assumption:** This method allocates final notes by aligning raw scores from top to bottom and allocating performance notes on the basis of some predetermined ratios. (For example, A for top 10%, B for the later 20%, C for the later 40% and so on)
2. **Evaluation according to the distribution intervals:** When the raw score distributions are examined, some gaps (discontinuities) within the distributions may be observed. In this method, these gaps are assumed to indicate natural intervals of success levels. (For example in a series of raw scores starting as 100, 98, 97, 97, 80, 78, 76..., there is a significant gap between 97 and 80. According to this method, the first 4 scores (from 100 to 97) will receive the highest note and the rest will be determined in the similar manner. )
3. **Evaluation according to standard notes:** Each score is evaluated on the basis of the degree of deviation from the average score. This is the most widely used method employed by international tests (such as GRE, CEEB, ACT, SAT...etc) and by many universities, since it is the most advanced of all. Therefore we don't discuss the first two methods in this article.

Let's examine the relative evaluation system of Istanbul University as an example for evaluation methods based on standard notes.

### ***An Example Methodology (Istanbul University)***

The university applies a 3 step methodology for transforming raw (final) scores into notes (relative results) (Keskin and Ertan, 2001):

1. Calculating z standard scores of raw scores
2. Transforming z standard scores into T standard scores
3. Converting T standard scores into notes

Relative scores vary from 0 to 100 and notes vary from AA to EE (AA, AB, BB, BC...etc)

If  $X_i$  is raw score of any student I and  $\bar{X}$  is the average raw score, and  $S$  is the standard deviation of raw scores, then z standard score of each student can be calculated with Eq 1.

$$z = \frac{X - \bar{X}}{S}$$

**Eq 1 – z score**

With this operation, raw scores are converted to a comparable scale. However z standard scores includes negative values too, therefore it is required to convert the z standard scores into positive numbers by rescaling in order to avoid possible problems in the later steps. The resulting numbers are called T standard scores and are calculated with Eq 2.

$$T = 10 \times z + 50$$

**Eq 2 – T score (1)**

However, there is one problem with this methodology. The resulting T scores will be relatively low in classes where the average raw scores are relatively high when compared with the scores in classes with higher scores. For example, a student with a raw score of 75 in a class where the average raw score is about 50 will have a T score of 65-70 approximately. However, another student with a raw score of 90, may not even have a T score of 60 in a class where average score is approximately 80. Therefore an adjustment is required according to the class success level. The university uses Table 1 for this adjustment. This table has been arranged according to the systematic of Ebel (Ebel, 1974) and presents how T standard score boundaries can be arranged for different class success levels while converting T scores into notes.

**Table 1 – Converting T standard scores into relative notes**

Average Score of Class	Equivalent Relative Notes of T Standard Scores							
	AA (4)	BA (3.5)	BB (3)	CB (2.5)	CC (2)	DC (1.5)	DD (1)	F (0)
> 80 - ≤100	≥ 57	52 - 56.99	47 - 51.99	42 - 46.99	37 - 41.99	32 - 36.99	27 - 31.99	< 27
> 70 - ≤ 80	≥ 59	54 - 58.99	49 - 53.99	44 - 48.99	39 - 43.99	34 - 38.99	29 - 33.99	< 29
> 62.5 - ≤ 70	≥ 61	56 - 60.99	51 - 55.99	46 - 50.99	41 - 45.99	36 - 40.99	31 - 35.99	< 31
> 57.5 - ≤ 52.6	≥ 63	58 - 62.99	53 - 57.99	48 - 52.99	43 - 47.99	38 - 42.99	33 - 37.99	< 33
> 52.5 - ≤ 57.5	≥ 65	60 - 64.99	55 - 59.99	50 - 54.99	45 - 49.99	40 - 44.99	35 - 39.99	< 35
> 47.5 - ≤ 52.5	≥ 67	62 - 66.99	57 - 61.99	52 - 56.99	47 - 51.99	42 - 46.99	37 - 41.99	< 37
> 42.5 - ≤ 57.4	≥ 69	64 - 68.99	59 - 63.99	54 - 58.99	49 - 53.99	44 - 48.99	39 - 43.99	< 40
<42.5	≥ 71	66 - 70.99	61 - 65.99	56 - 60.99	51 - 55.99	46 - 50.99	41 - 45.99	< 43

This methodology has many exceptions and 2 alternative methods are proposed for different circumstances. For further information refer to Keskin and Ertan (2001).

### **An Example Methodology (OSYM)**

Another example methodology, which can be discussed here, is employed by OSYM (Student Selection and Placement Center for Higher Education in Turkey) in nationwide university examinations. (OSYM, 2005)

Any student who wishes to pursue a higher education degree in Turkey must enter a nationwide examination. Additional to the score received from this examination, each student receives another score, which is called AOBP, calculated according to the educational performance of the student in the pre-university period. The weighted combination of these two scores constitutes the final score of a student.

While calculating AOBPs (stands for “Weighted Score for Secondary School Success Level”), OSYM takes high school graduation scores of students into consideration. The same problem of evaluation in our Performance Management System appears in the case of OSYM, too. Distributions of graduation scores vary widely in each school. While graduation scores are quite high in some schools such as Science High Schools due to the fact that the students are highly qualified; the graduation scores are quite low in some schools (especially in the public schools) due to the fact that the student diversity is quite high and the student profiles on average are not very good. Although the student profiles of most private schools are not better than the public schools, the graduation scores are again very high due to the fact that they don’t want to rough up students and upset their families.

OSYM assumes each school as a group by itself and applies a methodology similar to the previous example:

1. Transforming raw scores into T standard scores
2. Transforming T scores into a fixed range distribution (T’ scores)
3. Weighing T’ scores by the school’s success level

The details of the operations are presented below. Merging Eq 1 and Eq 2, T sores are directly calculated by Eq 3.

$$T = 10 \times \frac{X - \bar{X}}{S} + 50$$

**Eq 3 – T score (2)**

Later T standard scores are transformed into a fixed range distribution in which the highest score is equal to 100 and the lowest score is equal to 50. (See Eq 4)

$$T' = 50 + \frac{50 \times (\text{Student's T standard score} - \text{Min T standard score})}{\text{Max T standard score} - \text{Min T standard score}}$$

**Eq 4 – T' score**

At the 3<sup>rd</sup> step, OSYM assumes that the average score received by students of that school in the nationwide examination is an indicator of the level of success of this school. Based on this assumption, OSYM transforms school success scores into a distribution in which the highest score is equal to 200 and the lowest score is equal to 100. The resulting number is called “success coefficient” (A) and it is calculated with the following formula:

$$A = 100 + \frac{100 \times (\text{School's success level} - \text{Min Success Level})}{\text{Max success level} - \text{Min success level}}$$

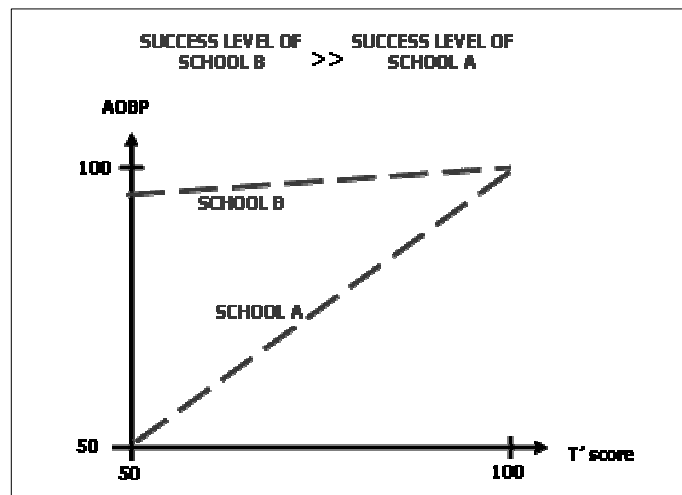
**Eq 5 – School success coefficient**

Later all students’ T’ scores are weighted by the school’s success coefficient according to the following formula and the AOBP is calculated.

$$AOBP = \frac{A \times (A + 225) \times (100 - T')}{125 \times (A + 160)} + 2 \times T' - 100$$

**Eq 6 – AOBP**

With this equation, top students in all schools receive the highest possible AOBP (100) and the other students’ AOBPs are rescaled according to the school success level as illustrated in Figure 7.



**Figure 7 – Interaction of school success coefficient and T' score**

## The Proposed Methodology

### *Introduction*

Due to the similarity of nature of the problems in evaluating educational performance in schools and evaluating performance measurement results of employees in corporations, the methodology proposed here is based on similar assumptions of the aforementioned methodologies. The proposed methodology consists of 3 main steps.

### *Step 1: Grouping*

At the very first step, it is required to detect people who are being affected from the same sources of variation and therefore have comparable scores. Here we have two rules:

- **The Same Origin Rule:** Since the most dominant variation source is the superintendent (as described in the preceding sections), we take organizational structure as the primary reference for identifying employee groups who must be treated together in the transformation process. Therefore, we group employees in such a way that each employee in a group is from the same organizational body, in other words, is evaluated by the same superintendent.
- **Population Level Rule:** Additionally, each group must catch up a population level, which theoretically must be at least 30, in order to secure statistical validity.

However, in most parts of the Tofas organization, it seems impossible to comply with both rules due to the complex organizational structure. Therefore, it is inevitable to stretch both rules for the sake of applicability. We have to merge some divisions since they have few employees. (Figure 8) In these cases, we have to assume that the superintendent of lower level superintendents preserves the balance among organizational groups reporting to his/her subordinates and assures that they are evaluated by these superintendents in the most “equal” way. In other words, we are going to treat performance scores of these employees as if they have been evaluated by the second level superintendent. We can include the lower level superintendents into the group too, as if they were not superintendents but ordinary employees evaluated by the same second level superintendent. (Of course, it is necessary to

inform the superintendents and employees in advance in order to let them spare time for formulizing and evaluating scorecards collaboratively) Additionally, we must also assume that groups of 5-10 can still secure statistical validity, since we are unable to reach 30 even if we stretch the “same origin” rule in some departments.

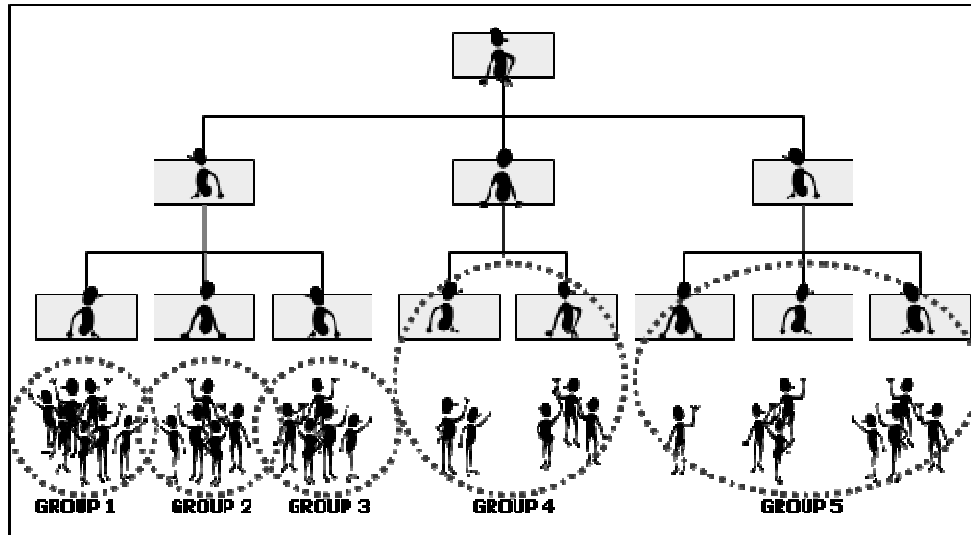


Figure 8 – An illustrated example for grouping

When we are done with the grouping at the bottom-line, we have to go up in the organizational hierarchy, and also group the management job family (MJF) staff in line with the same principles.

### **Step 2: Clustering**

After grouping employees, the next step is determining the set of groups to be compared with each other. While defining these sets (clusters), we face up a dilemma here. We have to decide which factor – micro or macro – is dominant in affecting performance measurement results. Let’s give an example to clarify the dilemma.

Think of a managerial control unit reporting to Purchasing Department’s head. This unit is responsible for carrying out managerial accounting and control tasks specific to Purchasing Department and work collaboratively with the main Budgeting & Control Department reporting to CFO. Actually, this unit is the decentralized part of the Budget and Control Department. Here we have two alternatives for clustering in this case. We can go in line with the administrative structure and put employees of Managerial Control Unit together with other Purchasing staff groups or we can take business relations and work commonality into

consideration and put this group into the cluster of Budget and Control Department. In the first case, we have to assume that Managerial Control employees in Purchasing are being affected from the same source of macro factors, or effects of those macro factors are negligible. Alternatively, in the second case, we have to assume that superintendent's effect is negligible or at least macro factors' effect is much more detrimental than the superintendent's effect. To resolve this dilemma, we have to decide which effect (superintendents' effect or macro factors' effect) is more detrimental.

This decision is especially important for clustering among management job family (MJF) employees. (As for MJF employees, it is inevitable to create cross departmental clusters since their population is lower and there is no alternative way to achieve a reasonable population level.)

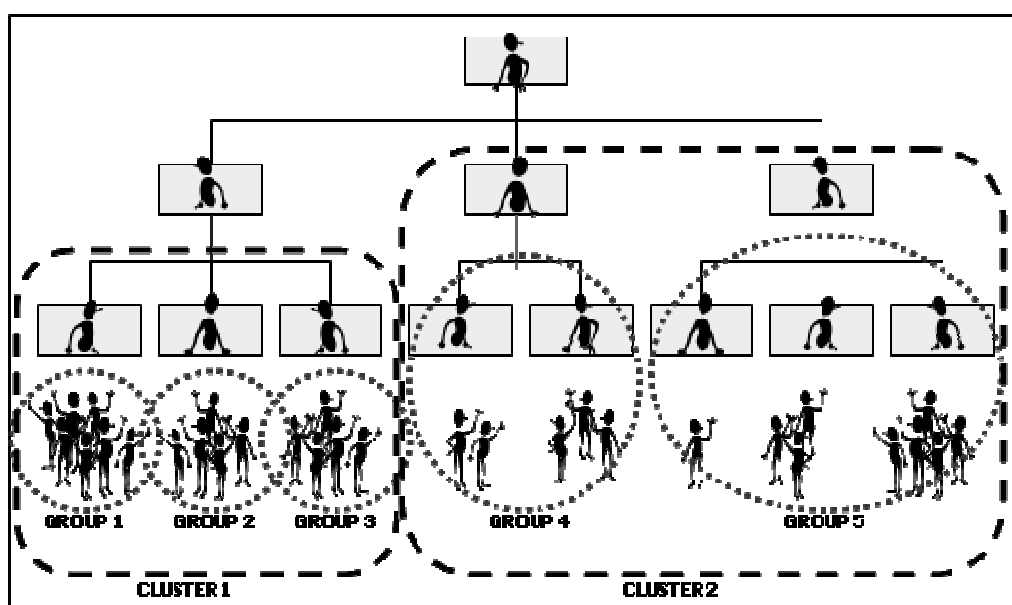


Figure 9 – An illustrated example for clustering

In Tofas, relying on our observations and experience, we concluded that the superintendent's effect is more detrimental than the macro factor's effect and therefore we made our clustering decisions based on the hierarchical structure.

### **Step 3: Transforming**

After grouping and clustering, the next step is to induce groups of performance scores into the same scale for comparison, and then compare these groups with each other. The process consists of four operations. The following abbreviations are used in the presented formulas:

AIP	: Absolute Individual Performance
AIG	: Average Individual Performance within the Group
SDG	: Standard Deviation within the Group
SIP	: Standard Individual Performance
Max(SIP) / Min(SIP)	: Max / Min SIP within the Group
RSIP	: Relative SIP
MAXARSIP	: Maximum Allowable RSIP
MINARSIP	: Minimum Allowable RSIP
AGP	: Absolute Group Performance
RGP	: Relative Group Performance
GSC	: Group Success Coefficient
Max(AGP) / Min(AGP)	: Max AGP among the Groups
WIP	: Weighted Individual Performance

### ***Operation 1: Standardizing***

Standardization is required to eliminate known sources of variation between groups and to transform the distribution of groups into normal distribution.

The first step of the standardization is pulling the means of performance distributions of groups to the same point so that the average performance score in each group becomes equal. In order to do this, group mean (AIG) is subtracted from all scores (AIP) within the group. With this operation, all information about mean location is lost and centered anomalies are obtained. Then the resulting scores in each group are divided by standard deviation of the group (SDG) in order to obtain normalized anomalies. This makes the data dimensionless. All information about location and scale is lost, so statistics based on standardized anomalies are left unaffected by any shifts or rescaling of the original data. (See Eq 7)

The resulting scores (SIP), which are called z-scores, are comparable with each other and represent mathematically the relative performance / success of all employees within the groups.

$$SIP = \frac{AIP - AIG}{SDG}$$

Eq 7 – SIP

NOTE [1]

For two reasons, skewness (symmetry) of score distributions is especially important to us. First reason is the methodology we employed at the last step of the transformation process (Operation 4: Classifying). As we assign notes sequentially according to the percentages of the pre-defined distribution, negatively skewed distributions become more advantageous or vice versa. Secondly, we do not need to worry about small group sizes when the sample population is approximately symmetric. (Thanks to the Central Limit Theorem (Tukey, 1977)) We can comfortably say that the score distributions become approximately normal for even small values of n when we apply the standardization process.

If skewness of a score distribution is high ( $>0.5$  or  $<-0.5$ ), then we must either look for constructing groups of at least 30 people or try to normalize the distribution. Normalization will make the data more normal (Gaussian). This will help reducing the skewness and the bias caused by outliers further and will transform data to better satisfy normality assumptions that are implicitly assumed by the statistical techniques used in this methodology. A simple normalization method may be taking median of AIP scores in replacement to AIG, while calculating SIP scores. If group populations are high enough to allow statistical validity, more complex methods can be used.

NOTE [2]

If SDG is too small (for example,  $< 1$  for scores over a scale of 0-100), then performance scores must be revised. Such a small standard deviation indicates something unusual in the measurement results.

### ***Operation 2: Making ranges equal***

At this step, SIP scores in each group are processed in such a way that score distributions of all groups have the same range. With this process, the lowest and highest scores in each

group will be equal and the other scores within the range will be redistributed accordingly. The resulting figures are called Relative Standard Individual Performance Scores (RSIP). (See Eq 8)

This step is required due to the final step employed for allocating performance notes. If this operation is not performed, then the larger groups, whose ranges are probably larger, will have higher amounts of extreme notes. (See Operation 4: Classifying)

$$RSIP = (MAXRSIP - MINARSIP) \times \frac{SIP - \text{Min}(SIP)}{\text{Max}(SIP) - \text{Min}(SIP)} + MINARSIP$$

**Eq 8 – RSIP**

NOTE [3]

When population of a group is too small ( $n < 4$ ) and that group can't be merged with any other group, it is not possible and also meaningless to apply the first 2 steps of the transformation process. A heuristic method for involving them into the process from the 3<sup>rd</sup> step may be calculating their RSIP by subtracting X (score of the least success level on the scale) from AIP and dividing the result with ([Maximum Possible AIP]-X); so that RSIP value of the maximum possible AIP score can't exceed 1.

$$RSIP = \frac{AIP - X}{(MaxPossibleAIP - X)}$$

**Eq 9 – RSIP (2)**

After that each RSIP value is multiplied by the related superintendent's success coefficient (RGP). If the related superintendent gets the highest score among others and hence the RGP is equal to 1, then Maximum Possible AIP will receive the maximum possible WIP.

NOTE [4]:

Values of MAXARSIP and MINARSIP are not important. Since working with negative values may be problematic, any value guaranteeing that all RSIP scores are positive can be used. We set MAXARSIP as 100 and MINARSIP as 50.

### **Operation 3: Relocating**

After the first two steps are complete, it can be assumed that all groups' distributions are converted to a comparable scale. There are two alternative routes at this point: You may not consider differentiating these two groups on the basis of group success and directly pass to the 4<sup>th</sup> step; or you may also incorporate the group success into evaluation process. We preferred to reflect the group success to some extent and we assumed that superintendents' scores represent the success levels of the groups reporting to those superintendents. However, here we have a restriction: The superintendents of all groups in a cluster must report to the same higher superintendent in order to assume that scores of superintendents are comparable with each other. Otherwise an alternative methodology must be applied. (It is possible to ask from the superintendent overseeing all groups in a cluster to rank the groups according to the perceived success level.)

At this operation, we compare the success of each group and shift resulting distributions according to their success levels by calculating a ratio linearly correlated with the performance of the groups' superintendents and multiplying it by every score within the groups. In order to do this, Absolute Group Performance (AGP) scores are transformed into a distribution in which the maximum AGP is equal to 1. (AGP of each group is the performance score of the superintendent to whom the group members report.) The transformed AGP scores are called Relative Group Performance (RGP) scores. They are calculated as follows:

$$RGP = \frac{AGP}{Max(AGP)}$$

**Eq 10 – RGP (1)**

Since we don't want the group success factor's impact to be very detrimental on the final results we impose a limitation here. If in the Eq 10, any group comes out to have a RGP score lower than GSC, then the following formula is used for calculating RGP of each group in a cluster:

$$RGP = (1-GSC) \times \frac{AGP - Min(AGP)}{Max(AGP) - Min(AGP)} + GSC$$

**Eq 11 – RGP (2)**

Determination of GSC is an entirely heuristic process and it is very much related with the principles of PMS and policies. By setting it at different levels, the effect of supervisors' success may be increased and decreased. With Eq 11, the minimum AGP becomes equal to the Group Success Coefficient (GSC) and the others are aligned between GSC and 1 accordingly.

After RGPs are calculated, each (RSIP) within a group are weighted by the success level (RGP) of that group; so that Weighted Individual Performance (WIP) scores are calculated. As a result, each group is favored against one another according to their success levels.

$$WIP = RGP \times RSIP$$

**Eq 12 – WIP**

NOTE [5]

If there are individuals (like secretaries) who can not be involved into a group, then it is not possible to apply the first 3 steps for them. Their scores must be involved into the process from the 4<sup>th</sup> step. A heuristic method for calculating their WIP may be subtracting X (score of the least success level on the scale) from AIP.

$$WIP = AIP - X$$

**Eq 13 – WIP (2)**

#### ***Operation 4: Classifying***

The last operation in the transformation process will be to assign performance notes. After relocating, all Weighted Individual Performance (WIP) scores in a cluster are ranked from highest to lowest and notes are assigned sequentially in accordance with the predefined ratios. (Top 3% are classified as “A”, the later 20%, 65% and 10% are classified as “B”, “C”, and “D” respectively, and the lowest 2% are classified as “E”)

#### ***An Alternative Methodology***

If it is not necessary to catch up with exact ratios of the predetermined distribution then an alternative way of transformation can be employed. After the standardization step is completed, the following process may be applied:

1. The standard scores (SIPs) are transformed into T standard scores (TIPs) by the help of Eq 14.

$$TIP = 10 \times SIP + 50$$

**Eq 14 – TIP score**

2. Later performance scores of superintendents (AGPs) are transformed into a scale over 100 (RGP') with Eq 15.

$$RGP' = 100 \times \frac{AGP}{150}$$

**Eq 15 – RGP'**

3. Performance notes are assigned by the help of Table 2, which has been derived from Table 1 according to the calculated TIP and RGP' values.

**Table 2 – Converting TIP scores into relative notes**

RGP'	TIP scores				
	A	B	C	D	F
> 80 - ≤100	≥ 57	47 - 56.99	32 - 46.99	27 - 31.99	< 27
> 70 - ≤ 80	≥ 59	49 - 58.99	34 - 48.99	29 - 33.99	< 29
> 62.5 - ≤ 70	≥ 61	51 - 60.99	36 - 50.99	31 - 35.99	< 31
> 57.5 - ≤ 52.6	≥ 63	53 - 62.99	38 - 52.99	33 - 37.99	< 33
> 52.5 - ≤ 57.5	≥ 65	55 - 64.99	40 - 54.99	35 - 39.99	< 35
> 47.5 - ≤ 52.5	≥ 67	57 - 66.99	42 - 56.99	37 - 41.99	< 37
> 42.5 - ≤ 57.4	≥ 69	59 - 68.99	44 - 58.99	39 - 43.99	< 40
<42.5	≥ 71	61 - 70.99	46 - 60.99	41 - 45.99	< 43

## Conclusion

There are many researchers and management commentators who have expressed doubts about the validity and reliability of the performance evaluation process. Some have even suggested that the process is so inherently flawed that it may be impossible to perfect it (Derven, 1990). However its crucial importance for organizational life can never be denied.

Although some of the assumptions lying beneath the proposed methodology are still being debated, the results have been found quite satisfactory by HR and company management and therefore the methodology has been in use since 2003 for approximately 900 white collar employees.

## References

1. **Derven, M.G.** (1990) The Paradox of Performance Appraisals, Personnel Journal, Vol 69
2. **Ebel, R. L.** (1965) Measuring Educational Achievement, Englewood Cliffs, N.J. Prentice Hall, Inc
3. **Ebel, R.L.** (1974) Marks and Marking Systems, IEEE Transactions on Education, V.
4. **Keskin, M. and Ertan, H.** (2001) Relative Evaluation System of Istanbul University, Istanbul University,
5. **Neely, A.** (2005) The Evolution of Performance Measurement Research, International Journal of Operations & Production Management, Vol 25
6. **OSYM** (2005) OSS Manual [Brochure]
7. **Tukey J. W.** (1977) Exploratory Data Analysis, Addison-Wesley,

## Indices

### **Tables**

Table 1 – Converting T standard scores into relative notes .....	14
Table 2 – Converting TIP scores into relative notes .....	25

### **Equations**

Eq 1 – z score .....	13
Eq 2 – T score (1).....	14
Eq 3 – T score (2).....	15
Eq 4 – T' score .....	16

Eq 5 – School success coefficient .....	16
Eq 6 – AOBP.....	16
Eq 7 – SIP .....	21
Eq 8 – RSIP.....	22
Eq 9 – RSIP (2).....	22
Eq 10 – RGP (1).....	23
Eq 11 – RGP (2).....	23
Eq 12 – WIP.....	24
Eq 13 – WIP (2).....	24
Eq 14 – TIP score.....	25
Eq 15 – RGP’ .....	25

### **Figures**

Figure 1 – The pre-defined distribution .....	7
Figure 2 – Score distributions of two different departments.....	7
Figure 3 – Different effects are at work in different phases.....	8
Figure 4 – Average target realization vs. average appraisal (discretion) score.....	9
Figure 5 – Correlation between target realization scores and appraisal (discretion) scores ..	10
Figure 6 – Standard deviation of target realization scores and appraisal (discretion) scores.	10
Figure 7 – Interaction of school success coefficient and T’ score .....	16
Figure 8 – An illustrated example for grouping.....	18
Figure 9 – An illustrated example for clustering.....	19