



|            |                                                                                                                                   |
|------------|-----------------------------------------------------------------------------------------------------------------------------------|
| Title      | Whether Tis Nobler to Normalize : Increasing Inter-evaluator Consistency in the Evaluation of Oral Communication Based Activities |
| Author(s)  | Inage, Itsuro; Lawn, Etsuko; Lawn, Murray                                                                                         |
| Citation   | 長崎大学教育学部紀要. 教科教育学, vol.49, pp.93-102; 2009                                                                                        |
| Issue Date | 2009-03-01                                                                                                                        |
| URL        | <a href="http://hdl.handle.net/10069/22122">http://hdl.handle.net/10069/22122</a>                                                 |
| Right      |                                                                                                                                   |

This document is downloaded at: 2012-05-17T04:01:19Z

# Whether Tis Nobler to Normalize

## — Increasing Inter-evaluator Consistency in the Evaluation of Oral Communication Based Activities —

稲毛 逸郎\*・ローン悦子\*\*・ローンマリー\*\*\*

(平成 20 年 10 月 30 日受理)

Itsuro INAGE\*, Etsuko LAWN\*\*, Murray LAWN\*\*\*

(Received, October 30, 2008)

### Abstract

When totaling evaluation scores in such as oral communication based activities in the case of multiple evaluators who have not been previously trained to given standards, a simple approach to increasing inter-evaluator consistency is the employment of normalization<sup>1</sup>.

### 1. Introduction

In regard to the evaluation of oral communication based activities, the provision of consistent and objective evaluation has been the subject of intense research by most major testing agencies that include the testing of oral communication ability (Inage & Lawn, 2005). The most commonly employed approach to increasing consistency and objectivity is to provide initial intense training, testing to ensure a given standard is met and then to provide periodic refresher courses thereafter. In the case of single evaluator based arrangements or globally objective evaluation, this is most likely unavoidable; however, in the case of multiple evaluators not requiring globally objective evaluation, the inherent redundancy of the multiple evaluators makes normalization feasible.

### 2. Background

In the authors' previous paper (Inage, Lawn & Lawn, 2007) a number of university students were interviewed by a native speaker of English, and the

---

\* 長崎大学教育学部教授・国際文化講座

\*\* 長崎大学長大教育機能開発センター非常勤講師

\*\*\* 長崎純心大学人文学部英語情報学講師

interviews were videotaped and evaluated by a wide variety of English teachers, both Japanese and native speakers of English. The evaluators were not previously trained to given standards, rather given simple categories to grade from. The evaluations were carried out by the individual teachers at their own convenience referring to the interviews via VHS tape, DVD or online as preferred. The inter-evaluator score variation<sup>2</sup> was significant. A careful analysis of the individual teacher's backgrounds, level of teaching (ranging from intermediate school to university) and native or non-native English speaker status showed very little conclusive correlation. The variation could only be categorized as "the variation of the individual," perhaps reflecting the character of the individual, the mood of the moment, etc. Normalization of the evaluator's results, however, showed very high inter-evaluator correlation. That is, while the resultant (raw) scores compared poorly, the inter-evaluator trends compared very closely. The observation of this high inter-evaluator correlation (score comparison after normalization) is the main focus of this paper. It is, therefore, proposed that normalization be considered in order to increase inter-evaluator consistency and objectivity, in the ad-hoc (irregular) evaluation of oral activities by multiple evaluators.

### 3. Normalization

Normalization cannot ensure global evaluation consistency as prior training will; however, it will simply ensure increased inter-evaluator consistency at a given time and place, which is ideally suited to ad-hoc events such as those requiring comparative precision only. This situation is perhaps most ideally suited to a competitive situation such as orally-based speech contests, where comparative placement of the participants is the main focus. In this case objective reference to some previously defined standard is of lesser interest. However, even in the case of evaluators having previous training, normalization of multiple evaluator results would surely be preferable to simple averaging.

### 4. Ensuring all evaluators have equal weighting

The mechanics of normalization focuses on underlying trends by negating the typical characteristics of the evaluator. For example, in the case that evaluator A gives participants an average of 80% with a standard deviation of 10%, compared to evaluator B giving the same participants an average of 70% and with a standard deviation of 5%. In regard to variation of average, simple averaging takes care of any variation. However, variation in standard deviation is not inherently compensated for. In contrast, the comparative standard deviation defines how much influence the individual evaluator has on the net result. Therefore, in the above case evaluator A has twice the influence on the final result compared to evaluator B. It is in this regard that the simple averaging of raw scores will result

in a bias toward the evaluator who scores more dynamically (score with a higher standard deviation).

## 5. Normalization exceptions

By normalizing the standard deviation, the above bias could be reduced (theoretically cancelled out); however, there may be some situations where by a specific score would perhaps best be removed from the normalization process. For example, if a participant fails to complete the required task, in the case of a speech or recitation contest, where perhaps a participant runs out of time or simply gives up halfway, the score must be lower compared to other participants though the degree of “penalty” may vary significantly between individual judges. It is, therefore, suggested that such cases be flagged as “exceptions” and be removed from the normalization calculation, and that the resulting grade be itself normalized. That is, this exception should be excluded from the normalization setup process; however, once the required normalization is calculated, it is recommended that the exception be subjected to the normalization in the same way other grades are. If such exceptions are not removed from the normalization calculation process, they may result in compensating for grading tendencies that should not be compensated for and result in negative impact on the fair grading of other participants, who comply with the requirements of the contest.

## 6. Resolving tied scores

In the case of evaluator grade totals being equal, (that is, ‘tied,’) the normalization process inherently tends to introduce a significant number of digits after the decimal point which provides clear placement for the participants. For the purposes of simplification, the number of digits shown after the decimal point has been kept to one in the examples used in this paper; however, for internal calculation purposes, the number of significant places has not been limited.

## 7. Analysis of a speech contest’s data

### 7.1 Simple averaging - example 1 (high variance in evaluator standard deviation)

The following example is used to show practically how normalization of data may be advantageous compared with simple averaging. The data is from an annual speech contest held at a private university in Nagasaki. The data represents two years of data, example 1 “year X” and example 2 “year Y.”

In the case of example 1 (year x), the variation in standard deviation between evaluators was significant, and as a result after normalization, the placement (e.g. 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> etc.) of participants (contestants) is greatly altered. In the case of example 2 (year y), the variation in standard deviation between evaluators was small, as a result after normalization, the placement of participants was almost

unchanged except for the resolution of a tied placement.

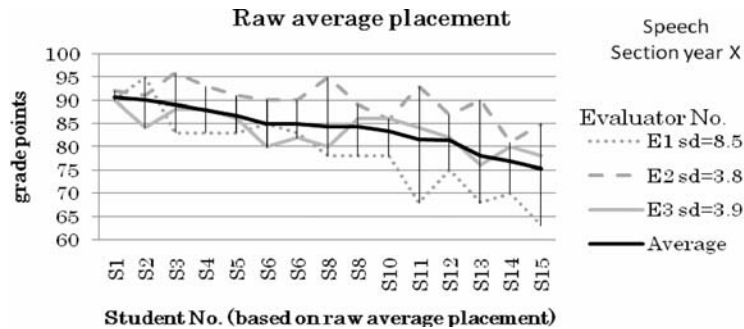


Figure 1. Placement based on simple averaging example 1 (year X)

Figure 1. shows the placement of participants using simple averaging. In this section, the participants are students labeled S1 etc. in Figure 1. The placement is shown by the S1 - S15 position on y axis in the graph; that is, Student 1 (S1) would receive 1<sup>st</sup> place based on simple (raw - original data as is) averaging and S2 2<sup>nd</sup> place, etc. It must be noted that S6 and S8 occur twice indicating tied scores. The grey lines on the graph indicate the grades given by three evaluators. The black line indicates the resultant average. Grade points are shown on the y axis. The respective standard deviation scores of the evaluators are Evaluator 1. (E1) 8.5 (grade points), E2 3.8 and E3 3.9. E1's standard deviation is of significance as it exceeds the combined standard deviations of E2 and E3; this implies that E1 has more than twice the influence of E2 and E3 in regard to the final result. In this particular speech contest, prior training of evaluators to given absolute grade points is not carried out; instead, simple categories are provided on which the evaluation is based. Thus, evaluators are free to grade as they see fit, although previous papers have noted grading correlation to be high under such a scheme. However, the standard deviation figure in effect defines the degree of weighting which a given evaluator has in regard to the final result (placement). In the case of inter-evaluator correlation being very high, such a variation in standard deviation will go unnoticed; however, in the case of variance in the evaluation of a specific participant, the evaluator exhibiting more variation in grading (higher standard deviation) will bias the result more than evaluators exhibiting less grading variation (lower standard deviation).

## 7.2 Result after normalization- example 1 (high variance in evaluator standard deviation)

Figure 2. is based on the same data as Figure 1. The evaluator's grading averages and standard deviations, however, have been normalized. The results in terms of placement are, most importantly, the re-ordering of 2<sup>nd</sup>, 3<sup>rd</sup> and 4<sup>th</sup>

places, and with the exception of 1<sup>st</sup> and 5<sup>th</sup> place remaining unchanged, significant reordering occurs in places 6 to 9. It should be noted that in this case the 6<sup>th</sup> and 8<sup>th</sup>, equal cases are resolved and given non-tied placements. The inherent complexity of the normalization calculations tends to resolve equal placement problems.

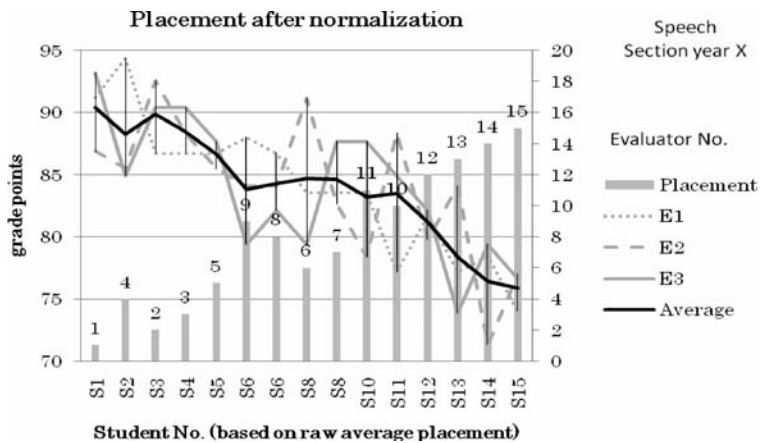


Figure 2. Placement based on normalization of evaluator standard deviations: example 1 (year X)

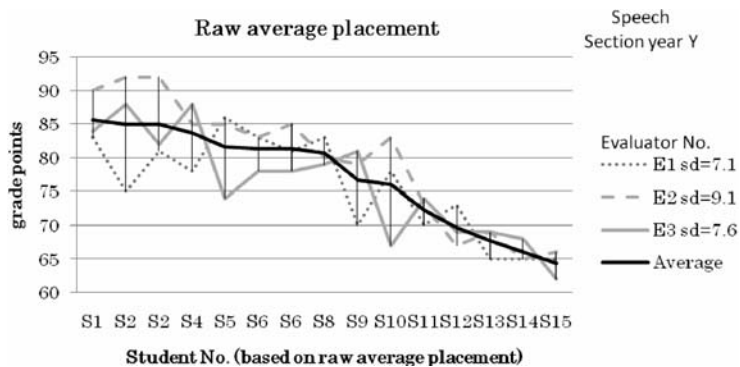


Figure 3. Placement based on simple averaging: example 2 (year Y)

### 7.3 Normalization of grade averages

While normalizing of grade averages has no effect on the final result (placement), it does give a more consistent look to the final grades. Normalization of the standard deviation, however, ensures that each evaluator has the same weighting in regard to the final result (placement). In addition to that, normalization of both the grade averages and standard deviation provides increased inter-evaluator consistency.

#### 7.4 Simple averaging - example 2 (low variance in evaluator standard deviation)

Figure 3. shows the second example of placement of participants using simple averaging. The point of variation compared to the data in Figure 1. is that the variation in standard deviation between evaluators is much less. E1 (evaluator 1) 7.1 (standard deviation), E2 9.1 and E3 7.6. representing standard deviation variations of 16% (E1 cf. E2) and 22% (E3 cf. E2) compared to example 1, where E1 8.5, E2 3.8 and E3 3.9 from Figure 1 represented standard deviation variations of 55% (E2 cf. E1) and 54% (E3 cf. E1).

#### 7.5 Result after normalization- example 1 (high variance in evaluator standard deviation)

The result after normalization in the case of closer evaluator standard deviations is that the change in placement is very small; in this case it is almost zero. The result is that the equal placement of 2<sup>nd</sup> is clearly given a non-tied result.

### 8. Practical implementation of the normalization of averages and standard deviations using a spreadsheet

The normalization of such as speech contest results can be carried out using a spreadsheet.

#### 8.1 Explanation of a sample normalization spreadsheet

Practical implementation of the normalization of averages and standard deviations using a spreadsheet based on Microsoft Excel is shown in Tables 1 and 2. In this spreadsheet standard deviations and averages are calculated in row 1. Row 2 shows the columns associated with Evaluator 1 (D-K). Row 3 indicates the column functions as follows from the left hand side. Column A shows raw placement, that is, natural raw score, for reference only. Column B shows the placement based on the normalizing process explained in this paper. Column C provides reference to the participant, e.g. name or number. Columns D to F in this case represent the grading subcategories. Column G gives the total raw score for evaluator 1. Columns H to J indicate the normalized subcategory grades. Column K shows the normalized total score for evaluator 1.

Columns D to K are repeated per evaluator. In this example, there are three evaluators. Rows 4 to 15 contain regular participant grades.

This spreadsheet provides for exceptions as discussed earlier in this paper, in rows 18 and 19 (section 5). It must be noted that the exceptions are not ranked and that the grades are not taken into account when calculating the individual evaluator's standard deviation.

In the case of the spreadsheet shown in Table 1, participant names are entered into column C and Evaluator 1's data entered into cells D, E and F, Evaluator 2's data into cells P, Q and R, etc. (not shown in Table 1).



Table 2 shows the spreadsheet's global calculation of all standard deviation averages and global averages, in this case for three evaluators. The spreadsheet has been set to show formulae for the purposes of this paper; usually it will show the results of each respective cells.

## 8.2 Spreadsheet calculations

The calculations made in the spreadsheet operate as follows. Firstly, each individual evaluator's score for each participant is totaled (G4 etc.). Next, the average (G1 etc.) and standard deviation (E1 etc.) of each evaluator are calculated. Then, the global average (AC1) and standard deviations (AE1) are evaluated. The individual evaluator's scores by participant are then calculated (K4 etc.) and the resulting rankings are recalculated (B4 etc.). As calculation of individual evaluator's scores by participant is a little complex, it is outlined in the following section.

## 8.3 Calculation of Cells K4 to 15 - individual evaluator's scores by participant

The mathematical calculations required to ascertain "individual evaluator's scores by participant" are shown as follows. Firstly, the variables are defined. The letter "n" represents the respective participants "1 to n" and evaluators "1 to n."

$P_{ngradeEn}$  Participant "n's" grade given by evaluator "n."

$E_{nav}$  Evaluator "n's" average grade.

$E_{nsd}$  Evaluator "n's" standard deviation

$E_{av}$  Average of all evaluation grades

$E_{sd}$  Average standard deviation of all evaluators

The calculation of the standard deviation and averaged normalized grades may be obtained step by step as follows:

$P_{nEndfa}$  Participant "n's" grade (by evaluator n) deviation from evaluator "n's" average

$$P_{nEndfa} = P_{ngradeEn} - E_{nav}$$

$P_{nEnsndnd}$  Participant "n's" grade (by evaluator n) standard deviation normalization deviation

$$P_{nEnsndnd} = P_{nEndfa} \times E_{sd} / E_{nsd}$$

$P_{nEnsnavn}$  Participant "n's" grade (by evaluator n) standard deviation and average normalized

$$P_{n\text{Ens}d\text{avn}} = P_{n\text{Ens}d\text{nd}} + E_{\text{av}}$$

By combining the above into a single calculation, the above  $P_{n\text{Ens}d\text{avn}}$  may be obtained by:

$$P_{n\text{Ens}d\text{avn}} = ((P_{n\text{grade}E_n} - E_{\text{av}}) \times E_{\text{sd}} / E_{\text{nsd}}) + E_{\text{av}}$$

The above calculation appears in cells K4 to L5 of the spreadsheet shown in Table 1.

## 9. Discussion

In the case of evaluation of speech by multiple non-trained evaluators, the process of normalization is often achieved by simply discussing results and verbally agreeing on placement of the important placements, typically 1<sup>st</sup> to 3<sup>rd</sup>. This informal process typically has a similar effect as this more formal normalization. However, it has been the authors' informal observation that dynamic graders, that is, evaluators who exhibit high standard deviations, tend to be equally as dynamic in presenting "their grades" in a more persuasive manner. This compares to say a more subtle grader who may be equally as subtle, that is, less persuasive in presenting their grades, if it comes to direct face to face discussions or confrontation regarding mutual placement of participants. It is in this regard that the more automated process may avoid this problem.

Regarding the normalized results, it is possible that the normalized results of "exceptions" previously discussed may give negative figures. In this case, the raw results may need to be used if feedback to the participant is required; however, a manual check in which the results are lower than non-exceptions would be recommended and appropriate adjustments would be made as required to ensure the presentation of realistic scores.

## 10. Conclusion

It has been proposed that normalization be applied to averages and standard deviations in order to provide increased inter-evaluator consistency in the evaluation of such as speech related activities by non-prior trained evaluators. The reason for the proposal is to ensure that each evaluator has the same level of influence on the result. While a similar normalization process may be attained by mutual discussions and score adjustment, the proposed automated process inherently provides the following advantages. The process is efficient, effective in ensuring fairness to all evaluators, avoids the need for discussions regarding the results, and it also provides inherent precision in inter-evaluator grading and usually resolves tied placements.

### Notes

1. “Normalization refers to the division of multiple sets of data by a common variable in order to negate that variable’s effect on the data, thus allowing underlying characteristics of the data sets to be compared.” (*Wikipedia* 2008) Specifically in the context of this paper, it is the inherent characteristics specific to the individual evaluator that we wish to negate in order to understand the underlying characteristics of the participant (the person being evaluated). In this paper, the term “normalization” is used to refer to the normalization of both score averages and more importantly the normalization of the standard deviation (a measure of the dispersion of a collection of values - (*Wikipedia* 2008)) unless otherwise specified.
2. Inter-evaluator score variation - the variation in scores between evaluators.

### References

- Arthur Hughes. 2006. *Testing for Language Teachers*. Cambridge: Cambridge University Press.
- Clifford Hill and Kate Parry. 1994. *From Testing to Assessment: English as an International Language*. London: Longman.
- Inage, I. and Lawn, E. 2006. “A preliminary Consideration of English Speaking Tests - Based on the Analysis of Current English Proficiency Tests -,” *Bulletin of Faculty of Education, Nagasaki University: Curriculum and Teaching*. No.46, pp. 137-151.
- Normalization (statistics), 2008/10/12, [http://en.wikipedia.org/wiki/Normalization\\_\(statistics\)](http://en.wikipedia.org/wiki/Normalization_(statistics))
- Standard deviation, 2008/10/12, [http://en.wikipedia.org/wiki/Standard\\_Deviation](http://en.wikipedia.org/wiki/Standard_Deviation)
- Inage, I., Lawn, E. and Lawn, M. 2007. “A Study of Native and Japanese Speakers of English Grading Tendencies of Speaking Ability - Based on the Analysis of Interview Evaluations and Background Questionnaire -,” *Bulletin of Faculty of Education, Nagasaki University: Curriculum and Teaching*. No.47, pp. 129-143.
- Lyle F. Bachman and Adrian S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.