Least square differences method for quantitative determination of rater bias

Author:

R. Clive Woods, Department of Electrical and Computer Engineering and Microelectronics Research Center, 2128 Coover Hall, Iowa State University, Ames, Iowa 50011, U.S.A.. cwoods@iastate.edu

Abstract *— In order to assess a large number of diverse projects as fairly and as rapidly as possible, a procedure often adopted is to use a panel consisting of a large number of assessors or raters, only a small number of whom assess each project. Since no single rater assesses all the projects, conscious or unconscious bias regarding overall standards by any rater will advantage or disadvantage the projects assessed by that particular rater. 'Rater bias' is the tendency for a project rater to return assessments that deviate from the required and established impartial standards so that the project(s) rated by that assessor are systematically advantaged or disadvantaged. The valuable criterion-referenced rating technique may be used to reduce rater bias, but assessing a project is still usually subjective and criterion-referencing does not entirely eliminate rater bias. Until recently only* ad-hoc *methods of determining rater bias were available. Perhaps the most direct was to distribute to the raters "standard examples" of project reports for assessment, having expected marks known to the overall investigator but not to the raters at the time of assessment; this method is impractical in evaluation procedures typically requiring several hours of work for a single complete assessment. Recently Woods [Woods, R.C., "Iterative processing algorithm to detect biases in assessments", IEEE Trans. Educ., 46(1), 2003, pp. 133–141] proposed a method of determining rater bias quantitatively from the complete set of genuine assessments, based upon a comparison of the 'paired assessments' of each rater with those other raters assessing the same projects. Independently Chan [Chan, K.L., "Statistical analysis of final year project marks in the computer engineering undergraduate program", IEEE Trans. Educ., 44(3), 2001, pp. 258–261] proposed a method of evaluating rater bias using a commercial statistical package. In the present paper, a third method of determining rater bias quantitatively is examined. This method is based upon finding the summed square deviations of each mark from the corresponding paired assessments, and minimizing this summed square deviation by adjustment of the means and standard deviations of each assessor in turn. The method has the advantage of placing the determination of rater bias on a more rigorous mathematical platform than previously, at the expense of more difficult data processing and slower numerical convergence. The algorithm is presented in detail, together with the results of applying it to the raw data used by both Woods and by Chan in presenting their original methods, so that the differences between the three techniques are clearly determined.*

Index Terms *— Assessment, grades, iterative algorithm, marks, peer-review, projects, proposals, rater bias.*

INTRODUCTION

A problem familiar to many persons working in higher education is that of assessing in a limited time a large number of student project reports. For example, these may be submitted in fulfillment of various requirements of many degree courses in all branches of engineering. If the number of projects is too large for assessment by one person in the time available, the usual procedure adopted is for a large number of assessors to be used; each project is assessed by only a small number of experts (typically two), and the assessors taken from the pool are nominally different for each project. A major example of where this procedure is often adopted is the assessment of student projects by academic examiners [1], as is now required as a part of most undergraduate degree courses. Other examples include the marking of essay-type examination questions where the essays are marked by a panel of examiners because of the workload involved, and also the peer-review or expert-evaluation of research proposals by a committee appointed by funds-awarding bodies [2].

In the present paper subsequently, for generality, a piece of work which is being assessed will be termed a "project"; a person undertaking evaluations will be termed a "rater"; the raters collectively form a body which will be termed the "panel"; and evaluation of a project produces a numerical "mark" awarded to that project.

A popular system, widely adopted, is for two raters drawn from the panel to be assigned to each project and to arrive at independent marks which are arithmetically averaged to give the final mark. This procedure relies upon each rater assessing the project using exactly the same standards that all the other raters adopt. The same standards must be used by all raters in order to assure that each individual project is neither advantaged nor disadvantaged depending upon which raters have been assigned to it (which assignment is sometimes quasi-random, and sometimes based upon considerations of the specific technical expertise area(s) of each rater). There are usually extra procedures adopted to try to ensure that each rater applies

the same standards, such as the use of detailed marking schemes [1,2] and requirements to produce a specified distribution of marks overall [2]. Usually, an extra rater from the panel is brought in as "moderator" if the original marks differ significantly, or if the original raters are unable to agree upon their marks. The definitions of "differ significantly" and "agree" are usually stated in terms of a certain number of points difference in the marks, which of course can vary from panel to panel. Minor variations upon this theme include:

- assigning other numbers of raters to each project;
- requiring raters to confer before the submission of agreed marks (or, alternatively, requiring them to submit initial marks *without* any discussion with other raters);
- either keeping raters unaware, or making them aware, (while undertaking the marking) of the identity of the other raters also assigned to their projects;
- using anonymous assessments, where the identity of each project's author is concealed from the raters assigned to assess it;
- recognizing differences in importance between different raters ("first" and "second" markers, where the "first" marker is required to be more actively involved in the project);
- requiring discussion between the raters in all cases after they have made their initial independent assessments, or alternatively requiring discussion only in the case of an initial significant disagreement; and
- either recording formally, or not, the original marks in cases where revised marks have resulted from a discussion between the raters concerned.

In some cases, variation of assessment standards is of little consequence. For example, in the early parts of a typical undergraduate course where all students undertake every one of a number of projects, it is of little concern if only one rater is used for all the submissions for one project. In this example, all the students will be assessed by this rater, and so all the students will have an equal advantage or disadvantage from that rater. Any mismatch of standards can then be handled, if necessary, by shifting all the marks as appropriate. On the other hand, small errors in project assessment by a small number of raters, as described in the previous paragraph, may have serious consequences. These can arise if the project mark forms a major part of the final degree result for a student, or if a funding panel is assessing projects and finds a large number of projects with similar marks clustered around the point which determines whether a project is to be funded or not. This is because it is difficult to determine whether a given rater habitually marks high or low (whether consciously or unconsciously), or whether the rater has been given an anomalously good or bad set of projects to assess. For example, a low average mark for any given rater may indicate either a poor batch of projects assessed by that rater, or a very low-marking rater, or a combination of these. If the assignment scheme is such that only one rater assesses each project, it is impossible to distinguish between these cases, and this scheme should, therefore, always be avoided. It also follows that a simple shifting of the marks returned by each rater, based upon the average of that rater's marks, is not necessarily a correct adjustment. As a consequence, in practice usually no adjustment is made.

There have been previous attempts [3,4] to modify examination marks in the case where each student is required to take only a certain number of examinations, but may choose these from a larger number of total examinations conducted. If one examiner is assigned to mark all the attempts at one examination, then biases can be detected by comparisons with the overall performance. However, such a procedure is inappropriate where each student submits only one (or a small number) of project reports or essays.

The technique of Correspondence Analysis has also been used to analyze examination mark data from individual questions in one examination [5]. This technique has additionally been applied to the case of several examinations, each taken by all students in a group [6]. None of these reports, however, discusses the case considered here, which is of enormous practical importance. The present situation is primarily characterized by each rater assessing only a relatively small number of projects.

Similarly, other techniques, such as meta-analysis and weighted graph optimization, are also not applicable in this case. For example, meta-analysis [7,8] is the process of combining information from a number of separate but essentially similar assessments by establishing a suitable statistical indicator that can be computed from each assessment. This is not a specific technique, but rather is a combination of other individual techniques. However, in the situation considered in the present paper each assessment is similar only to the other assessments of the same project. A further detailed method of handling such data has also been previously suggested [9], but this treatment permits only basic statistical processing of the marks.

In principle, distributing to the raters "standard examples" of project reports for assessment, having expected marks known to the investigator but not to the raters at the time of assessment, could be used as a method of detecting biases. This method has proven useful in other contexts [10], but it seems to be impractical in evaluation procedures typically involving reports of substantial length, requiring several hours of work for a complete assessment.

Recently, Chan [11] proposed a method of evaluating rater bias using scatter diagrams produced by a commercial statistical package. This work does not take into account, in a self-consistent manner, the variation in quality between individual sets of projects assessed by each rater. Nor does it take into account the possibility that certain raters award marks with a large standard deviation while others award marks with a low standard deviation.

Independently, Woods [12] proposed a method of determining rater bias quantitatively by analyzing the complete set of genuine assessments, based upon a comparison of the 'paired assessments' of each rater with those other raters assessing the same projects. This method takes into account both variation in quality between individual sets of projects assessed by each rater and also the possibility that certain raters award marks with a large standard deviation while others award marks with a low standard deviation.

In the present paper, a third method of determining rater bias quantitatively is examined. This algorithm allows statistical and quantitative determination of whether any raters appear to be making consistently harsh or bland assessments, as well as whether any projects appear to be given an undue advantage or disadvantage by their assignation of raters. This method is based upon finding the summed square deviations of each mark from the corresponding paired assessments, and minimizing this summed square deviation by adjustment of the means and standard deviations of each rater in turn. The method has the advantage over the previous method [12] of placing the determination of rater bias on a more rigorous mathematical platform than previously, but at the expense of more difficult data processing and programming, and rather slower numerical convergence. The complexity of the programming is a "one-time" problem that is soluble with care. Considerably slower convergence is not now a serious difficulty with the dataset size currently typical in Higher Education and funds-awarding bodies, provided a desktop computer is used having at least average performance. The algorithm is presented in detail below, together with the results of applying it to the raw data used by both Woods [12] and by Chan [11] in presenting their original methods, so that the differences between the three techniques may be clearly determined.

ALGORITHM FOR ASSESSING PROJECTS AND RATERS

If the projects are numbered from 1 upwards, and each project is assessed by a small number of raters drawn from a larger panel, then the conventional unmodified marks will appear as in Table 1. In this example there are two raters assigned to assess each project. In Table 1, the entry $M_{nm}^{(i)}$ denotes a mark of project n made by rater number m, and the superscript (i) denotes the iteration number: $i = 1$ signifies the original unmodified marks awarded. Therefore, this particular example table is to be interpreted as meaning that

- Project 1 was assessed by Rater A and Rater B;
- Project 2 was assessed by Rater C and Rater A;
- Project 3 was assessed by Rater D and Rater C;
- Project 4 was assessed by Rater B and Rater D; and
- the final mark reported for each project is the arithmetic average of the two raters' marks.

The model of the raters' assessment patterns assumed here is that any given rater will always give marks that differ from the ideal according to a shift of mean and a low or high standard deviation. It seems impossible to devise a prescriptive correction where an individual rater's evaluation differs significantly from the consensus in a *random* manner. If this possibility is suspected, then that rater's assessment criteria must be examined.

The fundamental principle of the procedure described below is based upon minimizing the summed squared differences between the marks awarded by each rater with the corresponding *paired assessments* [13]. The *paired assessments* are those marks awarded to the same projects by other raters [1]. The algorithm shifts the mean and varies the standard deviation of the marks awarded by each rater independently in turn so that, for each rater m, the summed square deviations of each mark from the corresponding paired assessments is minimized. The corrections made by this comparison must be iterated until convergence in order to be self-consistent. By using such a successive approximations process, a steady state will eventually be obtained. At this point, self-consistency in all the shifts and variations applied is achieved so that only random errors in marks should remain. It is assumed that each rater will assess a significant number of projects in total. Therefore, the paired assessments for each rater should be made by a representative sample of the panel, ideally approaching (or equaling) the number of projects assessed by that rater. This method differs significantly from the original algorithm presented in detail by Woods [12] where the algorithm adjusted the marks of each rater to minimize the *differences between the mean and standard deviation* of the corresponding paired assessments. In other words, this previous method was equivalent to matching as closely as possible the overall *distributions* of the marks of each rater with the corresponding paired assessments. This matching was achieved in that case by direct comparisons of means and standard deviations.

In more detail, the expression

$$F_m = \sum_y (p_y - q_y)^2 \tag{1}$$

is to be minimized, where

$$p_y = M_{ym}^{(i+1)} \tag{2}$$

is the mark awarded by rater m, and

$$q_y = M_{my}^{(i)} \tag{3}$$

is a paired assessment. This minimization is to be undertaken by shifting each rater's mean by Δ_m and simultaneously by multiplying each rater's standard deviation by Ω_m. No correction whatsoever corresponds to the parameter values $\Delta_m = 0$ and $\Omega_m = 1$. In practice, the next iteration of the marks awarded is computed using:

$$M_{nm}^{(i+1)} = \left(M_{nm}^{(i)} - \langle M_{xm}^{(i)} \rangle\right)\left(1 + \beta\left(\Omega_m - 1\right)\right) + \langle M_{xm}^{(i)} \rangle + \alpha\Delta_m \,, \tag{4}$$

where the average mark awarded by rater m is $\langle M_{xm}^{(1)} \rangle$ (where x is understood to indicate a mean taken over all the projects x to which rater m has been assigned), and α and β are iteration parameters that can be used to adjust the progress of the iterative process. The reason for introducing α and β here is that using positive values of α and/or β (less than unity) gives additional control over the iteration that may be exercised arbitrarily as required, and also gives significant protection against numerical instability as will be described below. The parameter α is associated with the correction of shifting the means of the marks. $\alpha = 0$ corresponds to no correction to the means, and a "normal" or full correction to the means corresponds to $\alpha = 1$. Intermediate values of α correspond to applying an intermediate proportion of the full correction to the shifting of the means at each iteration. Similarly, the parameter β is associated with the correction to the standard deviations of the marks. $\beta = 0$ corresponds to no correction to the standard deviations, and $\beta = 1$ to a full correction. Intermediate values of β correspond to applying an intermediate proportion of the full correction to the standard deviations at each iteration. Setting the values of α or β to zero allows the corrections to either the means or the standard deviations respectively to be suppressed for special purposes.

Minimizing F_m requires that Δ_m is the solution to

$$\frac{\partial F_m}{\partial \Delta_m} = 0 \,, \tag{5}$$

and also that Ω_m is the solution to

$$\frac{\partial F_m}{\partial \Omega_m} = 0 \,. \tag{6}$$

It may be shown by algebraic manipulation that

$$\Delta_m = \left(\frac{\Sigma q_y}{P}\right) - \left(\frac{\Sigma p_x}{N}\right) + \Omega_m \left[\left(\frac{\Sigma p_x}{N}\right) - \left(\frac{\Sigma p_x}{P}\right)\right], \tag{7}$$

where P is the total number of each rater's paired assessments (over y) and N is the total number of each rater's own assessments (over x), and that

$$\Omega_m = \frac{\sum_y \left[p_y - \left(\Sigma p_x / N\right)\right]\left[q_y - \left(\Sigma q_y / P\right)\right]}{\sum_y \left[p_y - \left(\Sigma p_x / N\right)\right]\left[p_y - \left(\Sigma p_y / P\right)\right]}, \tag{8}$$

so that the required correction may be obtained statistically. (Any occasional extra assessments by additional raters can also be included in these paired mean and standard deviation calculations.)

The iteration given by (4), (7) and (8) is applied in turn to the assessments of all projects n carried out by the raters m. If all projects are assessed the same number of times, then this procedure preserves the total number of marks and the overall standard deviation. In practice, the use of extra raters (or 'moderators') in cases of dispute (and for other reasons, such as quality control or monitoring of standards) causes the simple iteration not to preserve the overall mean or standard deviation. The correction formula:

$$M_{nm}^{(i+1')} = \langle M^{(1)} \rangle + (M_{nm}^{(i+1)} - \langle M^{(i+1)} \rangle)\sigma^{(1)} / \sigma^{(i+1)} \,, \tag{9}$$

will readjust the overall distribution, where $M_{nm}^{(i+1')}$ is the readjusted next iteration mark, $\langle M^{(1)} \rangle$ is the mean of all the original marks, $\langle M^{(i+1)} \rangle$ is the mean of all the un-readjusted next iteration marks, and $\sigma^{(1)}$ and $\sigma^{(i+1)}$ are the standard deviations of all original and next (un-readjusted) iteration marks respectively.

This iteration recipe is then repeated until convergence is obtained. It is useful to calculate the convergence factors given by the square root of $(\langle M_{xm}^{(i)} \rangle - \langle M_{xy}^{(i)} \rangle_m)^2$ averaged over all raters m (i.e. the root-mean-square or RMS differences of the individual raters' means and their paired means), and also the square root of $(\sigma_{xm}^{(i)} - (\sigma_{xy}^{(i)})_m)^2$ averaged over all raters m (i.e. the RMS differences of the individual raters' standard deviations and their paired standard deviations). These will reduce as the iteration proceeds, and their values can be examined to indicate a suitable point to terminate the iteration. At this point, the final adjusted mark for each project is the arithmetic average of the final iteration of the individual rater marks for that project.

DISCUSSION OF ALGORITHM PROPOSED

This iteration procedure is equivalent to a generalization of the multi-dimensional version of the well-known method of false position or *regula falsi* [14]. The standard method of false position is retrieved by setting $\alpha = \beta = 1$. However, in such a successive approximations calculation it is often found [14] that applying a correction that is too great can result in the problems of iterative "over-correction" and potentially numerical oscillation. Reducing the values of α and/or β slows the convergence but, more importantly, gives protection against numerical instability. Applying values of α and/or β that are too small leads to excessive computation time before convergence. Nevertheless, provided that numerical oscillation is avoided, *the final results are independent of the precise values of the parameters α and β* apart from changing the speed of the convergence [12]. Therefore, the criterion for choosing the values of α and β is to find the largest values of both parameters that, with all realistic datasets, consistently avoid the problems associated with numerical instability. In practice, as will be described below, with most realistic datasets and modern computers the elapsed computation time is no longer an important issue. Hence, there is no difficulty in practice in choosing suitable values for the convergence parameters α and β.

In the case of a rater whose marks are *perfectly* equi-distributed about the corrected values formed from the paired assessments, this scheme (using $\beta > 0$) has the effect of reducing this rater's standard deviation to zero. This also always occurs with raters who assess only one or two projects, since by definition their distributions are *perfectly* equi-distributed about the corrected values. It has a corresponding effect with reduced or imperfect equi-distribution. Iteration and convergence are much slower than the previous method [12], and programming it correctly is considerably more exacting and complex than programming the previous scheme [12]. Nevertheless, the use of a least-mean-squares calculation is mathematically and statistically more justifiable than the previous method [12], and so this method was pursued further here.

If only *one* project is assessed by any particular rater, then the standard deviation associated with that rater is zero, and a standard deviation correction is not possible for that rater. For any rater who assesses only *one or two* projects, the iteration of (4) will adjust those marks to be very close or equal to the paired assessments. For any rater who assesses a small number of projects (greater than two), the calculations of means and standard deviations for that rater will of course be subject to significant uncertainty, reducing the precision of the corrections in that case.

To illustrate this method using trivially simple data, consider the simple demonstration dataset shown in Table 2(a). This dataset is entirely fictitious and consists of marks awarded to three projects assessed twice each by a total of three raters, each rater undertaking two assessments. Rater A marks harshly such that his/her marks are always 5 less than the correct evaluation. Rater B always marks generously such that his/her marks are always 10 more than the correct evaluation. Finally, rater C always marks harshly such that his/her marks are always 5 less than the correct evaluation. The original marks awarded are printed in parentheses (), and the corresponding corrected values are those not enclosed in parentheses. Without knowing the rater biases displayed by the raters, the marks in parentheses show a wide disparity between the raters. But, applying the rater biases that are known in constructing the table, the marks are clearly seen to be correctable and self-consistent. The problem in practice is that the biases of the raters are obviously not known *a priori*; all that is available is a set of marks from which must be deduced the respective biases.

The biases can be deduced by comparing rater A's marks with rater B's mark for project #1 and with rater C's mark for project #3, together with other analogous comparisons, in a self-consistent manner. For example, comparing the two marks that rater A awarded with the corresponding paired marks, one would deduce from projects #1 and #3 that rater A marks too low by around (15 + 0)/2 = 7.5. Similarly, one would also deduce that rater B marks too high by around (15 + 15)/2 = 15, and that rater C marks too low by around (15 + 0)/2 = 7.5. Of course, if all the marks were adjusted precisely by these amounts then there would still be discrepancies in the marks, because the modifications alter the estimates of the biases. Therefore, the iteration must be repeated many times until a self-consistent set of adjustments has been obtained, as shown by the marks *not* contained in parentheses in Table 2(a). In this case, the self-consistent adjustments required are simply that rater A marks 5 too low, rater B marks 10 too high, and rater C marks 5 too low, which recover the biases assumed when this fictitious results table was originally constructed. The resulting average for each project is shown in the last column of Table 2(a), and the column headed "Δ" gives the changes in the quantity that is tabulated in the previous column. A summary of the final data is

listed by rater in Table 2(b). The final column of Table 2(b), headed "Paired Raters", indicates the number of raters with whom each rater has collaborated, and is included for consistency with the results tables given in the original paper [12].

In practice, the standard deviations of the marks awarded by each rater are also adjusted in a similar manner. For consistent presentation, in actual use the marks are further adjusted so that the overall mean and standard deviation of the entire set of marks are unchanged by the procedure.

When this procedure is used with actual data, the calculation is complicated by random assessment errors, which introduce some numerical noise. For accurate results relating to any given rater, that rater must assess a significant number of projects in total so that the statistical process will be meaningful. In practice, there may be some raters who assess only a small number of projects, but for completeness it is normal practice not to pre-select data and so all such raters are included in the analysis. Following the detailed statistical analysis described, there may be good reason to disregard or reduce the significance of results obtained for any raters who have assessed only a small number of projects. Of course, as in any statistical process, the more data that are available for any particular rater, then the more accurate will be the results relating to that rater. There is, of course, no sharp cut-off of numbers of projects below which accuracy is not obtained. Woods [12] discussed ways of improving the accuracy of the procedure by combining several datasets. Using datasets of the size often encountered in higher education (i.e., approximately 100 – 150 projects assessed by a panel of around 30 – 40 raters) then more weight will attach to the results from those raters assessing around 8 or more projects and less weight will attach to the results from those raters assessing around 3 or fewer projects, but these limits are not to be regarded as immovable. It should be recalled that under the conventional procedure (i.e., no adjustment whatsoever to the marks) it is common practice for each project to be assessed by just two raters whose marks are averaged (or, even worse, by only one rater), which procedure implies great weight attaching to just one or two assessments.

IMPLEMENTATION AND TESTING

The algorithm described above was programmed in high-level language, taking data input from a standard .CSV (Comma Separated Values) file for compatibility with widely-available spreadsheet programs. All of the detailed results described subsequently refer to the output of this program represented by (1) – (9) above. Using full-size datasets similar to those described below (Table 3), a typical computer (1.3GHz Pentium 4) calculated on the order of 140 iterations per second (elapsed time) for one dataset without storing the intermediate results from each iteration. A typical program implementation will process up to 400 projects assessed by 100 raters, and up to 3 assessments per project, though these limits may easily be increased on re-compilation.

Testing of the program proceeded with a number of artificial datasets. Some of these presented a small amount of trivial data, while others were full-size datasets (similar to Table 3) but artificially modified purely for the purpose of testing.

Modifications for testing included inserting fictitious problematic cases, such as raters giving very wildly disparate marks and also raters giving alternately high and low marks. Both of these special cases tested the numerical stability of the program under extreme conditions. In all cases, no problems were encountered in processing assessment data containing far more extreme assessment biases than would be encountered in practice, without evidence of assessment bias being quite clear from an inspection of a table of unprocessed marks. Tests were also undertaken to confirm that the ordering of the assessment records in the input data file did not affect the final results.

The program was further tested for reliably detecting abnormal marking patterns. This ability was investigated by manually subtracting a small value (e.g. 4%) from each mark awarded by one particular rater (chosen at random) in an otherwise genuine full-size dataset. On subsequently running the program, the program correctly modified the marks of this rater alone by adding to them the correct extra modification adjustment (4%) while leaving all the other modifications unchanged within 0.1%. Artificial modifications of this nature were detectable down to 0.1% (and probably lower), judged to be well below the practical accuracy limit of the basic assessments. As a further test, the distribution of one randomly chosen rater's marks was also artificially modified such that the standard deviation was doubled while keeping the mean the same. Again the program successfully identified this modification, by producing essentially the same set of modified marks as it did without this extra handicap, with minimal changes to the other modifications to the dataset.

One extremely important aspect of this program, also tested, is that the results should not be skewed by very high or very low marks, provided that these marks are accurate. The presence of a genuinely very high or very low quality project will significantly affect the means and standard deviations achieved by the raters assigned to assess it. Thus, at first sight, those raters *appear* to give abnormally high or low marks. However, if these raters agree on the project's high or low quality, then there is little effect upon the corrections applied to the marks by this algorithm. This aspect was tested by removing one genuinely very low quality project (assessed at below 10% by both raters, where the next highest project in that dataset was assessed at over 30%) from one of the real full-size datasets. This removal was found to make very little difference to the corrections for *any* of the raters in the panel. The means of the raters not assigned to this project changed by less than 0.1%, and their standard deviations by less than 0.5%; the raters assigned to this project clearly had large changes to their actual

means and standard deviations, but the adjustments calculated in their cases by the program were still virtually unchanged and within the same limits.

The program was finally applied to some genuine sets of project assessment data. The data in each genuine dataset were marks of around 130 student projects obtained from a single complete cohort of students at one particular time. The projects were all distinct and entirely separate from all those others represented in the same and the other datasets. All the raters were full-time faculty members (i.e., professional academics) in a major university department. Each project was normally assessed in an identical manner by two raters working independently. In both assessments, the starting data was produced objectively by marking each of eight or nine categories (as described previously [1]) to the nearest half-mark. These were summed to produce the final returned mark for each project on a scale from 0%, minimum, to 100%, maximum. Each dataset had an overall mean on the order of 55 ± 3% and an overall standard deviation on the order of 13 ± 2%. A large majority of panel members (around 30 raters) was common to all datasets so that recurring instances of abnormal marking patterns became obvious from an examination of all the results from the program.

Under certain conditions (only with the artificial datasets used in testing) the values of the iteration parameters $\alpha = \beta = 1.0$ were found to cause numerical instability. Hence, usually values of $\alpha = \beta = 0.5$ were used in processing the real datasets reported in these trials. Numerical instability was never observed with any datasets (genuine or artificial) for values of $\alpha = \beta \leq 0.5$. In practice, with the genuine datasets, the values of α and β were found to have little or no effect upon the values of the final results. This outcome is as expected, provided sufficient iterations were taken, and provided α and β were both greater than zero (giving no correction) and less than or equal to unity. With these values of α and β, convergence was obtained well within 500 iterations; usually, there was very little change after only 100 or 200 iterations. Use of these parameter values gave RMS differences of the individual raters' means and their paired means typically of less than 0.1%, and RMS differences of the individual raters' standard deviations and their paired standard deviations typically of less than 0.02%. Also after 500 iterations with genuine datasets, the changes in the raters' means for $\alpha = \beta = 0.5$ were all within 0.1% of those calculated using $\alpha = \beta = 1.0$. In addition, the calculated changes in the raters' standard deviations were all within 0.5%.

For direct comparison with the two methods previously described, presented here are the results of using this algorithm with the data originally used and presented by Woods [12] (hereinafter referred to as the "Data of Woods") and also the results of using this algorithm with the data originally used and presented by Chan [11] (hereinafter referred to as the "Data of Chan"). Previously these two authors have also examined the results of using their original methods to process the other's data [15], so that together with the present paper there are now published two separate datasets (obtained under similar conditions but by completely different personnel) with their corresponding rater biases evaluated by three different methods in total.

RESULTS AND DISCUSSION: DATA OF WOODS [12]

Table 3 shows the actual data output from the program using the Data of Woods [12]. The data in Table 3 were calculated using $\alpha = \beta = 0.5$. All values in parentheses represent unchanged (i.e. original and unmodified) data. The calculated results shown here were obtained after 892 iterations, although in practice there was little change identifiable after 500 iterations. To preserve anonymity, each project is here identified by an Arabic numeral, and the raters are identified by either one or two capital letters. The data table lists each project name, the two raters (Rater#1 and Rater#2) who assessed that project, the original marks (in parentheses), the adjusted marks, and the final average mark for each project. The column headed "Δ" is the difference between the original and adjusted average marks.

The full program output also includes (but, for clarity, not reproduced here) the basic program parameters; the adjustment given to the overall mean and standard deviation; the two convergence factors (i.e. the RMS difference between all the raters' means and their corresponding paired means, and the RMS difference between all the raters' standard deviations and their corresponding paired standard deviations); and the value of the standard deviation of the average correction applied to each project mark (giving a measure of how consistent were the original uncorrected marks).

Finally, the program provides summary data for each rater, as shown in Table 4 calculated from the same dataset as in Table 3. Again, original unmodified values are shown in parentheses, and the same one or two letter codes are used to identify the assessing raters. This table lists each rater's original and final adjusted means and standard deviations. The columns headed "Δ" indicate the changes of value of the mean mark or standard deviation respectively, calculated for each rater. The column headed "Total" is the total number of assessments carried out by each rater. Finally, the column headed "Paired Raters" is the number of different raters who were paired with that rater. Also included in the full listing (but, for clarity, omitted here) is the total number of paired assessments. The last three values are particularly useful in checking the results to confirm that the program has correctly allied all the marks to the correct raters. This information guards against the possibility of, for example, a misspelt rater's name in a project entry in the input file (which, if not corrected, would erroneously generate two rater calculations for one person).

In the original paper of Woods [12], projects 3, 28, 30, and 58 were identified as poor projects requiring large adjustments. The present program agrees with the identification of project 28 as requiring large correction upwards. Also, projects 114 and 127 were identified as requiring large corrections to their marks which were probably more accurate than those of poor projects. The present program agrees with these findings. Using the data described, most raters in practice returned marks that required adjustments to their average of less than ~3%. However, some required larger adjustments, sometimes up to on the order of 2 or 3 times this value within any one particular dataset. Also, most raters returned a personal standard deviation on the order of 10%. Generally speaking, adjustments to their average of consistently greater than 5% (shown by typically three or four raters in a dataset of this size) were interpreted as cause for concern regarding assessments.

RESULTS AND DISCUSSION: DATA OF CHAN [11]

In the Data of Chan [11], there are a number of projects where both raters awarded a mark of 0. This arises where no project submission is made by that student, and it seems reasonable to exclude these projects from the algorithm since, by definition, awarding these marks requires no judgment and the marks do not need adjusting. Also, a large number of raters assessed only one or two projects (whereas there are fewer of these in the Data of Woods [12]); as noted above, their standard deviations would be reduced to zero by applying a convergence parameter $\beta > 0$ to adjust their standard deviations. More seriously, an examination of the Data of Chan [11] shows that, in fact, this dataset consists of the following distinct and self-contained subsets, defined by the raters linked by paired assessments:

- Subset #1, 21 projects: Raters #{2, 4, 8, 14, 16, 17, 19, 25, 32, 41, 45}
- Subset #2A, 7 projects: Raters #{1, 24, 29, 33, 35, 36, 38}
- Subset #2B, 2 projects: Raters #{11, 15, 37}
- Subset #3, 79 projects: Raters #{5, 6, 9, 10, 12, 13, 18, 20, 21, 22, 23, 26, 28, 30, 34}
- Subset #4, 3 projects: Raters #{27, 39, 40, 46}
- Subset #5, 4 projects: Raters #{3, 44}
- Subset #6, 1 project: Raters #{31, 42}
- Subset #7, 1 project: Raters #{7, 43}

There are no rater pairings between any members of differing subsets; all rater pairings are between raters who are members of the same subset. Subsets 2A and 2B would be combined if one project (#103) assessed at 0 by both raters #11 and #29 were re-included. The situation is illustrated graphically in the "pairings diagrams" (Figs. 1 and 2 for the Data of Woods [12] and Chan [11] respectively) showing the individual pairings between raters. In these diagrams, each line indicates a pairing between the lower numbered (or lettered) rater to the higher of the two. The diagrams are drawn irrespective of which rater took primary responsibility for the assessment (the "first marker" or "project supervisor") and which had secondary responsibility (the "second marker" or "second assessor") since in the algorithms presented here and also previously by Woods [12] the two are treated as equally important. (The raters for each project are, of course, also treated as equally important in conventional results processing where the marks from separate raters are simply averaged arithmetically to obtain the overall mark for each project.) The line thickness drawn is proportional to the number of projects shared between each pairing of raters; the thinnest lines indicate that just one project was shared between the two raters concerned. The line color indicates the subset of each pairing.

Since the Data of Chan [11] is composed of these eight self-contained subsets, this total dataset is in fact ill-conditioned as far as the present algorithm is concerned. As each subset is self-contained it is possible in principle to apply the algorithm separately to each subset, but applying it in full form to the complete dataset reduces many raters' standard deviations to zero since the program will 'latch' onto one subset and correct that at the expense of the other subsets. Moreover, apart from subset #3, the subsets each contain too few projects for reliable statistical accuracy. There are two possible approaches here. Firstly, the algorithm may be applied only to those self-contained subsets of the data that have sufficient projects for reliable statistics. Alternatively, the standard deviation parameter β may be set to zero so that no iterative adjustment of standard deviations occurs; in this latter case, the program may be used to assess all the projects. In practice, often there is only limited information contained in adjustments to the standard deviations anyway [12].

Generally, such pathological pairings of raters should always be avoided since it is entirely feasible for one subset of raters to use standards completely different from another subset. Since there is no linkage or comparison made between the subsets, there can be no possible moderation of this situation. There are no such problems with the Data of Woods [12] since in that dataset each rater is linked to all the others either directly or indirectly through other pairings.

Tables 5 and 6 show the result of processing all the Data of Chan [11] using iteration parameters $\alpha = 0.5$ and $\beta = 0.0$ so that the standard deviations were retained nominally unchanged. (In fact all are adjusted slightly by the same factor, as the

program also automatically maintains the same overall standard deviation.) 713 iterations were used, again with little change observed after 500 iterations. Tables 3, 4, 5, and 6 are to be compared directly with the results reported in [11], [12] and [15].

The alternative approach is illustrated in Tables 7 and 8 which show the results of using the algorithm to process only subset #3 of the Data of Chan [11]. Here, 2093 iterations were used but convergence is actually much faster than this number might indicate, since there are only about half the total number of projects involved in this calculation. The original programs found that raters 21, 36, and 44 marked significantly too low; this is confirmed by the overall processing (Table 6) but not by the subset processing (Table 8) which leads to speculation that the previous conclusions (which of course used the complete dataset) may have been misled by the pathological nature of this dataset.

CONCLUSIONS

Generally, these algorithmic methods have proven extremely useful in identifying immediately any raters who habitually give high or low marks. They also identify those with a systematic narrow or wide assessment range. In addition, the program output readily identified those projects assessed by raters with strong assessment biases in the same direction. These were subject to large corrections by the program. All of this information is available from an *inspection* of the program results, without the necessity of actually *formally* reporting or recording the modified marks.

For quantitative evaluation of rater bias from raw assessment data, a third technique is now available in addition to the two methods previously published. This method is considerably more complex to program and is slower than the other methods, but is more mathematically rigorous. The overall results are similar to those obtained by other methods, though the detailed results differ slightly as the three methods are not exactly equivalent.

The Data of Chan [11] actually consists of eight self-contained subsets of assessment data, each subset containing rater pairings only amongst its members. This type of pathological dataset may be examined either in full using the present algorithm by annulling the standard deviation adjustment, or by each individual subset alone where the subset contains enough projects for statistically meaningful conclusions to be drawn. In undertaking assessments of this type it is, however, preferable to avoid such pathological pairings of raters.

REFERENCES

[1] Allison, J., and Benson, F.A., "Undergraduate projects and their assessment", *IEE Proc. A,* Vol. 130, 1983, pp. 402-419

[2] *Guide for evaluators, Marie Curie fellowships (individual and host)*, European Commission: Brussels, 2000 (also available at URL: ftp://ftp.cordis.lu/pub/improving/docs/mcf_guide_evaluators.pdf)

[3] Seed, N.L., *The examination algorithm*, unpublished report, University of Sheffield, 1998

[4] Radley, D.E., *Modification of final-year examination marks*, unpublished report, University of Sheffield, 1986

[5] Greenacre, M.J., *Theory and applications of correspondence analysis*, Academic Press: London, 1984, pp. 271-280

[6] Murtagh, F., "Verifying examination results: a general approach", *Special Interest Group on Comp. Sci. Education Bulletin*, Vol. 14, 1982, pp. 2-11

[7] Moore, D.S., and McCabe, G.P., *Introduction to the practice of statistics*, 3rd. ed., W.H. Freeman & Co.: New York, 1998, pp. 632-633

[8] Wolf, F., *Meta-analysis, quantitative methods for research synthesis*, Sage Publications: Newbury Park, 1986

[9] Teo, C.Y., and Ho, D.J., "A systematic approach to the implementation of final year projects in an electrical engineering undergraduate course", *IEEE Trans. Educ.*, Vol. 41, 1998, pp. 25-30

[10] Engelhard, G., Davis, M., and Hansche, L., "Evaluating the accuracy of judgments obtained from item review committees", *Appl. Measurement in Educ.*, Vol. 12, 1999, pp. 199-210

[11] Chan, K.L., "Statistical analysis of final year project marks in the computer engineering undergraduate program", *IEEE Trans. Educ.*, Vol. 44(3), 2001, pp. 258-261

[12] Woods, R.C., "Iterative processing algorithm to detect biases in assessments", *IEEE Trans. Educ.*, Vol. 46(1), 2003, pp. 133-141

[13] McSweeney, B., and Bingen, G., private communication, 1999

[14] Isaacson, E., and Keller, H.B., *Analysis of numerical methods*, Wiley: New York, 1966, p. 99

[15] Woods R.C. & Chan K.L.: 'Comparison of two quantitative methods of determining rater bias', *ASEE J. Eng. Ed.*, Vol. 92, 2003, pp. 295-306

FIGURES

FIGURE 1

PAIRINGS DIAGRAM FOR THE DATA OF WOODS [12].

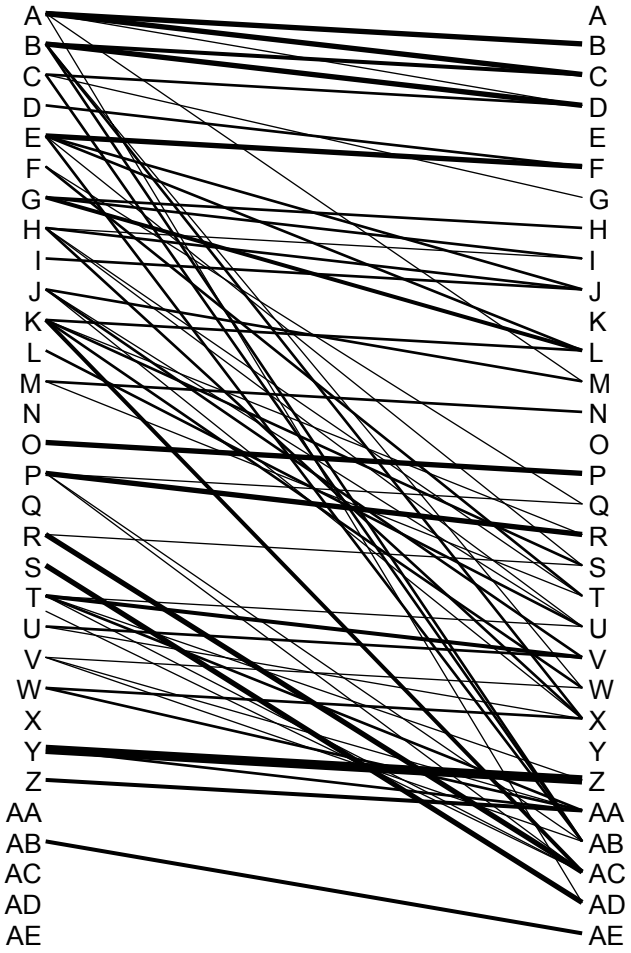
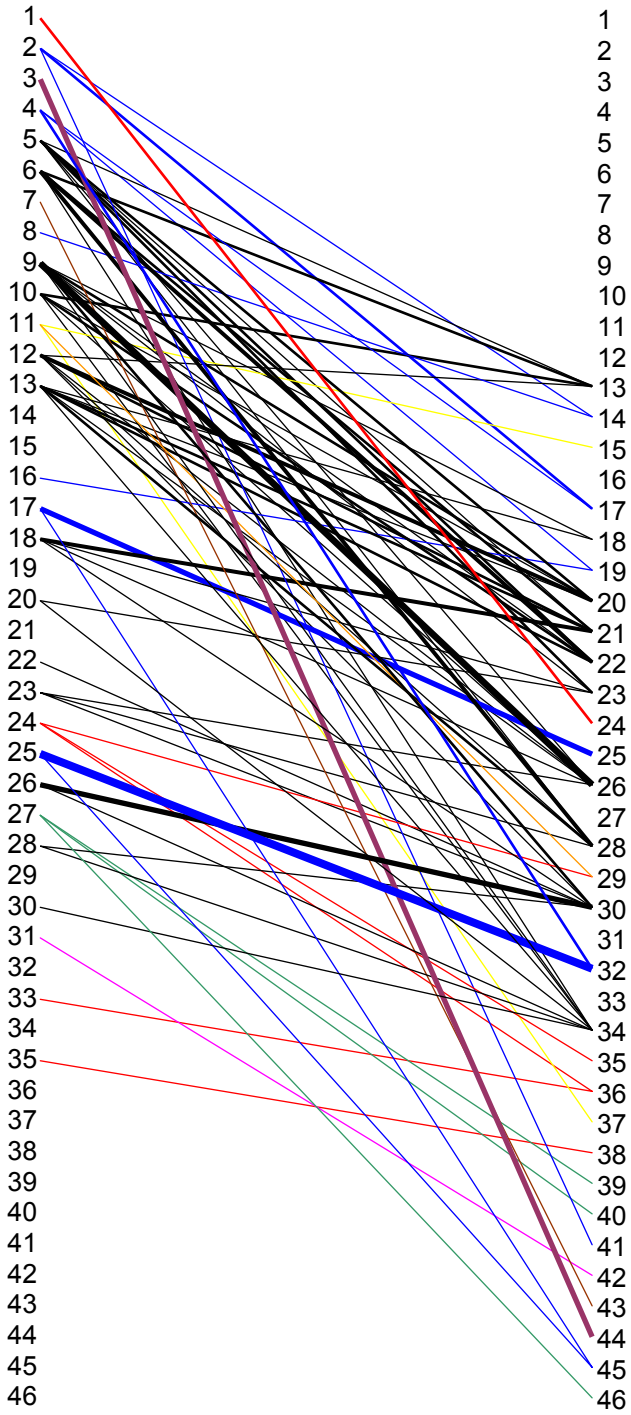


FIGURE 2

PAIRINGS DIAGRAM FOR THE DATA OF CHAN [11]. (COLOR KEY: SUBSET #1 = BLUE; SUBSET #2A = RED; SUBSET #2B = YELLOW; SUBSET #3 = BLACK; SUBSET #4 = GREEN; SUBSET #5 = MAGENTA; SUBSET #6 = PURPLE; SUBSET #7 = BROWN.)



TABLES

TABLE 1

TYPICAL RESULTS TABLE WHEN TWO RATERS EVALUATE EACH PROJECT.

Name		Mark 1		Mark 2	Average mark
Project 1	Rater A	$M_{11}^{(1)}$	Rater B	$M_{12}^{(1)}$	$(M_{11}^{(1)} + M_{12}^{(1)})/2$
Project 2	Rater C	$M_{23}^{(1)}$	Rater A	$M_{21}^{(1)}$	$(M_{23}^{(1)} + M_{21}^{(1)})/2$
Project 3	Rater D	$M_{34}^{(1)}$	Rater C	$M_{33}^{(1)}$	$(M_{34}^{(1)} + M_{33}^{(1)})/2$
Project 4	Rater B	$M_{42}^{(1)}$	Rater D	$M_{44}^{(1)}$	$(M_{42}^{(1)} + M_{44}^{(1)})/2$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮

TABLE 2

FICTITIOUS DATA USED TO ILLUSTRATE THE PRESENT METHOD.

(*a*) *Data listed by project:*

Project	Rater#1	<--Marks-->	Rater#2	<--Marks-->	<-Averages>	Δ
#1	A	(75.0) 80.0	B	(90.0) 80.0	(82.5) 80.0	-2.5
#2	B	(55.0) 45.0	C	(40.0) 45.0	(47.5) 45.0	-2.5
#3	C	(60.0) 65.0	A	(60.0) 65.0	(60.0) 65.0	5.0

(*b*) *Summary of data sorted by rater:*

Rater	<--Means-->	Δ	<-Std Dev->	Δ	Total	Paired Raters
A	(67.5) 72.5	5.0	(7.5) 7.5	0.0	2	2
B	(72.5) 62.5	-10.0	(17.5) 17.5	0.0	2	2
C	(50.0) 55.0	5.0	(10.0) 10.0	0.0	2	2

TABLE 3

OUTPUT FROM THE PROGRAM USING DATA OF WOODS [12].

Project	Rater#1	<--Marks-->	Rater#2	<--Marks-->	<-Averages>	Δ
#1	A	(49.0) 44.7	B	(48.0) 45.2	(48.5) 45.0	-3.5
#2	B	(70.0) 65.4	C	(66.5) 66.3	(68.3) 65.9	-2.4
#3	D	(21.0) 21.8	B	(24.0) 23.2	(22.5) 22.5	0.0
#4	E	(41.0) 46.9	F	(50.0) 50.4	(45.5) 48.7	3.2
#5	G	(50.0) 52.8	H	(54.0) 52.7	(52.0) 52.7	0.7
#6	I	(51.0) 52.4	J	(51.0) 49.0	(51.0) 50.7	-0.3
#7	K	(38.0) 39.5	L	(36.0) 39.2	(37.0) 39.3	2.3
#8	L	(57.0) 57.4	K	(54.0) 56.6	(55.5) 57.0	1.5
#9	M	(68.0) 63.7	N	(66.0) 63.9	(67.0) 63.8	-3.2
#10	O	(55.0) 58.7	P	(51.0) 57.2	(53.0) 58.0	5.0
#11	B	(63.0) 59.0	A	(62.0) 60.0	(62.5) 59.5	-3.0
#12	F	(37.0) 42.1	Q	(41.0) 41.8	(39.0) 41.9	2.9
#13	P	(46.0) 50.4	O	(48.0) 49.1	(47.0) 49.8	2.8
#14	R	(56.0) 61.2	S	(58.0) 59.2	(57.0) 60.2	3.2
#15	D	(58.0) 54.6	A	(58.0) 55.3	(58.0) 54.9	-3.1
#16	C	(62.0) 61.9	G	(60.0) 62.3	(61.0) 62.1	1.1
#17	T	(69.0) 65.2	F	(65.0) 60.1	(67.0) 62.6	-4.4
#18	U	(46.5) 46.9	V	(49.0) 46.2	(47.8) 46.6	-1.2
#19	V	(55.0) 52.7	T	(48.5) 48.5	(51.8) 50.6	-1.1
#20	T	(58.0) 56.3	V	(55.0) 52.7	(56.5) 54.5	-2.0
#21	E	(37.0) 44.7	F	(37.0) 42.1	(37.0) 43.4	6.4
#22	B	(57.0) 53.5	D	(62.0) 58.1	(59.5) 55.8	-3.7
#23	K	(39.0) 40.5	W	(41.0) 43.8	(40.0) 42.2	2.2

#24	L	(52.0) 53.1	E	(60.0) 57.5	(56.0) 55.3	-0.7
#25	X	(63.0) 60.5	U	(60.0) 60.2	(61.5) 60.3	-1.2
#26	M	(70.0) 65.8	J	(68.0) 66.1	(69.0) 66.0	-3.0
#27	X	(38.0) 41.2	H	(39.0) 41.2	(38.5) 41.2	2.7
#28	Y	(21.0) 41.8	Z	(24.0) 41.2	(22.5) 41.5	19.0
#29	T	(46.5) 46.9	AA	(42.0) 46.8	(44.3) 46.8	2.6
#30	J	(36.0) 34.0	M	(39.0) 33.4	(37.5) 33.7	-3.8
#31	AB	(57.5) 53.9	C	(56.0) 55.9	(56.8) 54.9	-1.9
#32	L	(49.0) 50.5	U	(49.0) 49.4	(49.0) 49.9	0.9
#33	Z	(73.0) 63.3	AA	(67.0) 62.0	(70.0) 62.7	-7.3
#34	Z	(54.0) 54.8	AA	(52.0) 52.9	(53.0) 53.8	0.8
#35	F	(51.0) 51.1	E	(47.0) 50.3	(49.0) 50.7	1.7
#36	H	(50.0) 49.6	T	(52.5) 51.8	(51.3) 50.7	-0.5
#37	H	(65.0) 61.0	X	(66.0) 62.8	(65.5) 61.9	-3.6
#38	AA	(51.0) 52.3	Y	(53.0) 54.9	(52.0) 53.6	1.6
#39	I	(51.0) 52.4	G	(48.0) 50.9	(49.5) 51.7	2.2
#40	T	(63.5) 60.7	AC	(60.0) 62.1	(61.8) 61.4	-0.3
#41	A	(74.0) 74.0	C	(72.0) 71.8	(73.0) 72.9	-0.1
#42	AC	(70.0) 74.4	K	(69.0) 72.6	(69.5) 73.5	4.0
#43	E	(40.0) 46.4	L	(46.0) 47.8	(43.0) 47.1	4.1
#44	X	(47.0) 48.1	B	(50.0) 47.1	(48.5) 47.6	-0.9
#45	M	(54.0) 49.1	N	(52.0) 49.0	(53.0) 49.0	-4.0
#46	T	(50.0) 49.8	U	(53.0) 53.3	(51.5) 51.5	0.0
#47	P	(56.0) 64.0	R	(55.0) 59.8	(55.5) 61.9	6.4
#48	G	(38.0) 41.4	I	(35.0) 40.9	(36.5) 41.2	4.7
#49	R	(51.0) 54.1	P	(50.0) 55.8	(50.5) 55.0	4.5
#50	W	(68.0) 65.3	K	(67.0) 70.5	(67.5) 67.9	0.4
#51	Y	(67.0) 60.6	Z	(61.0) 57.9	(64.0) 59.3	-4.7
#52	AC	(74.0) 79.4	R	(72.0) 84.0	(73.0) 81.7	8.7
#53	R	(47.0) 48.4	AC	(50.0) 49.8	(48.5) 49.1	0.6
#54	B	(51.0) 48.0	C	(50.0) 50.0	(50.5) 49.0	-1.5
#55	AA	(67.0) 62.0	Y	(66.0) 60.2	(66.5) 61.1	-5.4
#56	J	(61.0) 59.1	I	(59.0) 58.2	(60.0) 58.6	-1.4
#57	K	(49.0) 51.2	R	(49.0) 51.2	(49.0) 51.2	2.2
#58	A	(35.0) 28.3	B	(29.0) 27.8	(32.0) 28.1	-3.9
#59	AA	(53.0) 53.5	Z	(47.0) 51.6	(50.0) 52.5	2.5
#60	AA	(27.0) 37.6	T	(38.0) 40.0	(32.5) 38.8	6.3
#61	Y	(48.0) 52.8	Z	(52.0) 53.9	(50.0) 53.3	3.3
#62	S	(60.0) 61.5	K	(58.0) 60.9	(59.0) 61.2	2.2
#63	C	(58.0) 57.9	AB	(61.0) 56.3	(59.5) 57.1	-2.4
#64	E	(75.0) 65.8	J	(65.0) 63.1	(70.0) 64.4	-5.6
#65	F	(61.0) 57.5	D	(61.0) 57.2	(61.0) 57.4	-3.6
#66	P	(73.0) 87.2	Q	(74.0) 87.8	(73.5) 87.5	14.0
#67	A	(66.0) 64.7	C	(66.0) 65.8	(66.0) 65.2	-0.8
#68	L	(54.0) 54.8	G	(52.0) 54.7	(53.0) 54.7	1.7
#69	B	(62.0) 58.0	W	(66.0) 63.7	(64.0) 60.9	-3.1
#70	D	(58.0) 54.6	C	(50.0) 50.0	(54.0) 52.3	-1.7
#71	P	(61.0) 70.9	R	(62.0) 69.8	(61.5) 70.3	8.8
#72	J	(48.5) 46.5	E	(43.5) 48.3	(46.0) 47.4	1.4
#73	A	(68.0) 67.0	AD	(66.0) 68.9	(67.0) 68.0	1.0
#74	Z	(67.0) 60.6	T	(66.0) 62.8	(66.5) 61.7	-4.8
#75	Y	(52.0) 54.5	Z	(60.0) 57.5	(56.0) 56.0	-0.0
#76	AB	(63.0) 57.7	B	(60.0) 56.2	(61.5) 56.9	-4.6
#77	M	(64.0) 59.6	A	(62.0) 60.0	(63.0) 59.8	-3.2
#78	B	(60.0) 56.2	AB	(56.0) 52.8	(58.0) 54.5	-3.5
#79	J	(68.0) 66.1	V	(65.0) 63.5	(66.5) 64.8	-1.7
#80	B	(54.0) 50.7	A	(54.0) 50.6	(54.0) 50.6	-3.4
#81	K	(37.0) 38.4	AC	(39.0) 36.2	(38.0) 37.3	-0.7
#82	H	(48.0) 48.1	U	(45.0) 45.5	(46.5) 46.8	0.3
#83	W	(45.0) 47.0	AA	(45.0) 48.6	(45.0) 47.8	2.8
#84	V	(48.0) 45.1	AB	(46.0) 45.9	(47.0) 45.5	-1.5
#85	W	(72.0) 68.4	X	(74.0) 69.0	(73.0) 68.7	-4.3

#86	V	(56.0)	53.8	W	(54.0)	54.1	(55.0) 54.0	-1.0
#87	AC	(46.0)	44.8	P	(41.0)	43.6	(43.5) 44.2	0.7
#88	V	(63.0)	61.3	U	(59.5)	59.7	(61.3) 60.5	-0.7
#89	K	(56.5)	59.2	AC	(58.0)	59.6	(57.3) 59.4	2.2
#90	S	(64.0)	66.0	AD	(64.0)	67.0	(64.0) 66.5	2.5
#91	AC	(67.0)	70.7	R	(64.0)	72.6	(65.5) 71.7	6.2
#92	B	(66.0)	61.7	X	(64.0)	61.3	(65.0) 61.5	-3.5
#93	U	(64.0)	64.1	L	(66.0)	65.2	(65.0) 64.7	-0.3
#94	O	(69.0)	77.9	P	(66.0)	77.7	(67.5) 77.8	10.3
#95	S	(73.0)	76.2	AD	(75.0)	77.8	(74.0) 77.0	3.0
#96	B	(59.0)	55.3	D	(55.0)	51.9	(57.0) 53.6	-3.4
#97	Z	(58.0)	56.6	Y	(55.0)	55.7	(56.5) 56.1	-0.4
#98	X	(61.0)	58.9	W	(62.0)	60.5	(61.5) 59.7	-1.8
#99	G	(59.0)	61.3	L	(61.0)	60.9	(60.0) 61.1	1.1
#100	AB	(54.0)	51.4	AE	(54.0)	53.3	(54.0) 52.4	-1.6
#101	X	(54.0)	53.5	J	(57.0)	55.1	(55.5) 54.3	-1.2
#102	D	(69.0)	64.3	B	(75.0)	70.0	(72.0) 67.1	-4.9
#103	AA	(43.0)	47.4	W	(41.0)	43.8	(42.0) 45.6	3.6
#104	AA	(62.0)	59.0	V	(62.0)	60.3	(62.0) 59.6	-2.4
#105	C	(48.0)	48.0	B	(52.0)	48.9	(50.0) 48.4	-1.6
#106	S	(52.0)	52.4	AD	(48.0)	51.2	(50.0) 51.8	1.8
#107	C	(50.0)	50.0	A	(55.5)	52.3	(52.8) 51.1	-1.6
#108	E	(63.0)	59.1	AB	(66.0)	59.8	(64.5) 59.4	-5.1
#109	S	(78.0)	81.9	AD	(77.0)	79.7	(77.5) 80.8	3.3
#110	AC	(66.0)	69.5	T	(68.0)	64.4	(67.0) 67.0	-0.0
#111	G	(52.0)	54.7	L	(54.0)	54.8	(53.0) 54.7	1.7
#112	F	(65.0)	60.1	T	(62.5)	59.9	(63.8) 60.0	-3.8
#113	H	(47.0)	47.3	G	(44.0)	47.1	(45.5) 47.2	1.7
#114	Y	(75.0)	63.9	Z	(74.0)	63.8	(74.5) 63.8	-10.7
#115	AE	(68.0)	60.6	AB	(65.0)	59.1	(66.5) 59.8	-6.7
#116	AE	(60.0)	56.4	AB	(66.0)	59.8	(63.0) 58.1	-4.9
#117	T	(50.5)	50.2	V	(53.0)	50.5	(51.8) 50.4	-1.4
#118	M	(67.0)	62.7	T	(64.5)	61.6	(65.8) 62.1	-3.6
#119	H	(47.0)	47.3	J	(51.0)	49.0	(49.0) 48.2	-0.8
#120	AB	(68.0)	61.2	E	(65.0)	60.2	(66.5) 60.7	-5.8
#121	O	(78.0)	90.3	P	(75.0)	90.0	(76.5) 90.1	13.6
#122	AB	(47.0)	46.6	P	(41.0)	43.6	(44.0) 45.1	1.1
#123	C	(52.0)	51.9	A	(52.0)	48.2	(52.0) 50.1	-1.9
#124	R	(40.0)	38.4	AC	(39.0)	36.2	(39.5) 37.3	-2.2
#125	F	(73.0)	65.2	E	(68.0)	61.9	(70.5) 63.5	-7.0
#126	D	(73.0)	67.9	F	(76.0)	67.1	(74.5) 67.5	-7.0
#127	Z	(70.0)	62.0	Y	(72.0)	62.7	(71.0) 62.3	-8.7
#128	R	(87.0)	105.5	P	(87.0)	106.3	(87.0) 105.9	18.9
#129	J	(53.0)	51.0	H	(49.0)	48.9	(51.0) 50.0	-1.0
#130	C	(52.0)	51.9	D	(56.0)	52.8	(54.0) 52.4	-1.6
#131	U	(57.5)	57.7	J	(57.5)	55.6	(57.5) 56.7	-0.8
#132	E	(62.0)	58.6	S	(58.5)	59.8	(60.3) 59.2	-1.1
#133	V	(62.0)	60.3	J	(62.0)	60.1	(62.0) 60.2	-1.8
#134	K	(70.5)	74.2	S	(72.0)	75.1	(71.3) 74.6	3.4
#135	H	(58.0)	55.7	I	(50.0)	51.7	(54.0) 53.7	-0.3

TABLE 4
SUMMARY DATA SORTED BY RATER, FOR THE SAME DATASET SHOWN IN TABLE 3 (DATA OF WOODS [12]).

Rater	<--Means-->		Δ	<-Std Dev->		Δ	Total	Paired Raters
A	(57.8)	55.0	-2.8	(10.1)	11.8	1.7	11	5
B	(55.0)	51.6	-3.4	(12.9)	11.8	-1.1	16	6
C	(56.9)	56.8	-0.1	(7.6)	7.6	-0.1	12	5
D	(57.0)	53.7	-3.3	(13.9)	12.3	-1.6	9	4
E	(54.7)	54.5	-0.2	(12.6)	7.0	-5.6	11	5

F	(57.2)	55.1	-2.2	(13.5)	8.7	-4.8	9	4
G	(50.4)	53.2	2.8	(6.8)	6.5	-0.3	8	4
H	(50.8)	50.2	-0.6	(7.0)	5.3	-1.7	9	6
I	(49.2)	51.1	1.9	(7.8)	5.6	-2.2	5	3
J	(56.5)	54.6	-1.9	(8.9)	8.9	0.0	12	7
K	(53.8)	56.4	2.6	(12.2)	13.0	0.8	10	5
L	(52.8)	53.7	1.0	(8.2)	7.1	-1.1	9	4
M	(60.3)	55.7	-4.6	(10.8)	11.3	0.5	6	4
N	(59.0)	56.4	-2.6	(7.0)	7.5	0.5	2	1
O	(62.5)	69.0	6.5	(11.7)	16.1	4.4	4	1
P	(58.8)	67.9	9.1	(14.3)	19.6	5.2	11	5
Q	(57.5)	64.8	7.3	(16.5)	23.0	6.5	2	2
R	(58.3)	64.5	6.2	(12.9)	18.5	5.5	10	4
S	(64.4)	66.5	2.1	(8.4)	9.5	1.1	8	4
T	(56.7)	55.2	-1.5	(9.3)	7.6	-1.7	13	8
U	(54.3)	54.6	0.3	(6.5)	6.4	-0.1	8	6
V	(56.8)	54.6	-2.2	(5.7)	6.1	0.5	10	6
W	(56.1)	55.8	-0.3	(11.8)	9.4	-2.4	8	5
X	(58.4)	56.9	-1.5	(10.8)	8.3	-2.4	8	5
Y	(56.6)	56.3	-0.2	(15.5)	6.3	-9.1	9	2
Z	(58.2)	56.6	-1.5	(13.6)	6.2	-7.5	11	3
AA	(50.9)	52.2	1.3	(11.8)	7.2	-4.6	10	5
AB	(59.0)	54.9	-4.1	(7.3)	5.1	-2.2	11	6
AC	(56.9)	58.3	1.4	(12.1)	15.0	2.8	10	4
AD	(66.0)	68.9	2.9	(10.3)	10.1	-0.2	5	2
AE	(60.7)	56.8	-3.9	(5.7)	3.0	-2.7	3	1

TABLE 5
OUTPUT FROM THE PROGRAM USING COMPLETE DATA OF CHAN [11].

Project	Rater#1	<--Marks-->		Rater#2	<--Marks-->		<-Averages>		Δ
#1	14	(72.0)	71.1	2	(70.0)	71.1	(71.0)	71.1	0.1
#2	17	(70.0)	73.0	2	(74.0)	75.0	(72.0)	74.0	2.0
#3	17	(0.0)	4.5	2	(0.0)	2.6	(0.0)	3.5	3.5
#4	19	(66.0)	66.4	4	(68.0)	66.4	(67.0)	66.4	-0.6
#5	24	(72.0)	75.6	36	(65.0)	75.6	(68.5)	75.6	7.1
#6	33	(78.0)	78.5	36	(68.0)	78.5	(73.0)	78.5	5.5
#7	13	(58.0)	56.3	5	(60.0)	63.2	(59.0)	59.8	0.8
#8	20	(80.0)	79.2	5	(76.0)	78.9	(78.0)	79.0	1.0
#9	21	(75.0)	78.6	5	(70.0)	73.0	(72.5)	75.8	3.3
#10	22	(78.0)	76.7	5	(75.0)	77.9	(76.5)	77.3	0.8
#11	22	(73.0)	71.8	5	(68.0)	71.0	(70.5)	71.4	0.9
#12	26	(77.0)	76.5	5	(76.0)	78.9	(76.5)	77.7	1.2
#13	28	(74.0)	71.9	5	(60.0)	63.2	(67.0)	67.6	0.6
#14	11	(70.0)	68.7	37	(73.0)	68.7	(71.5)	68.7	-2.8
#15	13	(60.0)	58.3	6	(55.0)	58.9	(57.5)	58.6	1.1
#16	21	(83.0)	86.5	6	(73.0)	76.5	(78.0)	81.5	3.5
#17	21	(77.0)	80.6	6	(72.0)	75.5	(74.5)	78.0	3.5
#18	22	(80.0)	78.6	6	(65.0)	68.6	(72.5)	73.6	1.1
#19	22	(83.0)	81.6	6	(72.0)	75.5	(77.5)	78.5	1.0
#20	28	(70.0)	68.0	6	(60.0)	63.8	(65.0)	65.9	0.9
#21	28	(79.0)	76.8	6	(65.0)	68.6	(72.0)	72.7	0.7
#22	34	(68.0)	60.1	6	(66.0)	69.6	(67.0)	64.8	-2.2
#23	18	(67.0)	65.9	9	(68.0)	70.5	(67.5)	68.2	0.7
#24	20	(60.0)	59.6	9	(65.0)	67.5	(62.5)	63.6	1.1
#25	22	(70.0)	68.8	9	(59.0)	61.6	(64.5)	65.2	0.7
#26	26	(71.0)	70.6	9	(72.0)	74.4	(71.5)	72.5	1.0
#27	26	(75.0)	74.5	9	(71.0)	73.4	(73.0)	74.0	1.0
#28	26	(73.0)	72.6	9	(69.0)	71.4	(71.0)	72.0	1.0
#29	30	(55.0)	54.2	9	(54.0)	56.8	(54.5)	55.5	1.0

#30	35	(30.0)	22.7	38	(35.0)	22.7	(32.5)	22.7	-9.8
#31	13	(65.0)	63.1	10	(68.0)	64.9	(66.5)	64.0	-2.5
#32	13	(30.0)	28.9	10	(30.0)	27.7	(30.0)	28.3	-1.7
#33	22	(55.0)	54.2	10	(60.0)	57.1	(57.5)	55.6	-1.9
#34	26	(61.0)	60.8	10	(62.0)	59.0	(61.5)	59.9	-1.6
#35	34	(80.0)	71.8	10	(70.0)	66.9	(75.0)	69.3	-5.7
#36	15	(70.0)	63.8	11	(65.0)	63.8	(67.5)	63.8	-3.7
#37	27	(56.0)	56.4	39	(63.0)	56.4	(59.5)	56.4	-3.1
#38	20	(62.0)	61.6	12	(52.0)	52.6	(57.0)	57.1	0.1
#39	21	(62.0)	65.9	12	(58.0)	58.5	(60.0)	62.2	2.2
#40	21	(72.0)	75.7	12	(73.0)	73.2	(72.5)	74.4	1.9
#41	26	(73.0)	72.6	12	(68.0)	68.3	(70.5)	70.4	-0.1
#42	30	(0.0)	0.4	12	(0.0)	1.7	(0.0)	1.1	1.1
#43	6	(60.0)	63.8	13	(60.0)	58.3	(60.0)	61.0	1.0
#44	12	(72.0)	72.2	13	(68.0)	66.1	(70.0)	69.1	-0.9
#45	22	(70.0)	68.8	13	(69.0)	67.1	(69.5)	67.9	-1.6
#46	22	(80.0)	78.6	13	(70.0)	68.0	(75.0)	73.3	-1.7
#47	26	(70.0)	69.7	13	(58.0)	56.3	(64.0)	63.0	-1.0
#48	28	(50.0)	48.4	13	(65.0)	63.1	(57.5)	55.8	-1.7
#49	8	(65.0)	59.3	14	(60.0)	59.3	(62.5)	59.3	-3.2
#50	4	(75.0)	73.3	17	(72.0)	75.0	(73.5)	74.1	0.6
#51	25	(56.0)	55.5	17	(54.0)	57.4	(55.0)	56.4	1.4
#52	25	(60.0)	59.4	17	(50.0)	53.5	(55.0)	56.4	1.4
#53	25	(60.0)	59.4	17	(50.0)	53.5	(55.0)	56.4	1.4
#54	25	(56.0)	55.5	17	(60.0)	63.2	(58.0)	59.4	1.4
#55	13	(77.0)	74.9	18	(68.0)	66.9	(72.5)	70.9	-1.6
#56	21	(50.0)	54.2	18	(59.0)	58.0	(54.5)	56.1	1.6
#57	21	(30.0)	34.6	18	(35.0)	34.6	(32.5)	34.6	2.1
#58	23	(65.0)	67.0	18	(65.0)	63.9	(65.0)	65.4	0.4
#59	26	(51.0)	51.1	18	(62.0)	61.0	(56.5)	56.0	-0.5
#60	16	(43.0)	36.1	19	(35.0)	36.1	(39.0)	36.1	-2.9
#61	27	(70.0)	70.1	40	(70.0)	70.1	(70.0)	70.1	0.1
#62	5	(60.0)	63.2	20	(68.0)	67.4	(64.0)	65.3	1.3
#63	9	(74.0)	76.3	20	(68.0)	67.4	(71.0)	71.9	0.9
#64	10	(67.0)	63.9	20	(60.0)	59.6	(63.5)	61.8	-1.7
#65	12	(74.0)	74.2	20	(65.0)	64.5	(69.5)	69.3	-0.2
#66	12	(60.0)	60.5	20	(60.0)	59.6	(60.0)	60.0	0.0
#67	13	(66.0)	64.1	20	(66.0)	65.5	(66.0)	64.8	-1.2
#68	23	(58.0)	60.1	20	(65.0)	64.5	(61.5)	62.3	0.8
#69	13	(72.0)	70.0	21	(45.0)	49.3	(58.5)	59.6	1.1
#70	13	(67.0)	65.1	21	(53.0)	57.1	(60.0)	61.1	1.1
#71	18	(73.0)	71.7	21	(70.0)	73.7	(71.5)	72.7	1.2
#72	6	(40.0)	44.2	22	(20.0)	19.9	(30.0)	32.0	2.0
#73	10	(56.0)	53.2	22	(55.0)	54.2	(55.5)	53.7	-1.8
#74	30	(69.0)	67.9	22	(60.0)	59.0	(64.5)	63.5	-1.0
#75	5	(63.0)	66.1	23	(66.0)	67.9	(64.5)	67.0	2.5
#76	5	(62.0)	65.2	23	(52.0)	54.2	(57.0)	59.7	2.7
#77	9	(49.0)	51.9	23	(60.0)	62.1	(54.5)	57.0	2.5
#78	26	(60.0)	59.9	23	(60.0)	62.1	(60.0)	61.0	1.0
#79	28	(76.0)	73.9	23	(70.0)	71.9	(73.0)	72.9	-0.1
#80	30	(53.0)	52.2	23	(50.0)	52.3	(51.5)	52.3	0.8
#81	2	(65.0)	66.2	41	(60.0)	66.2	(62.5)	66.2	3.7
#82	1	(74.0)	75.1	24	(72.0)	75.6	(73.0)	75.3	2.3
#83	1	(58.0)	59.4	24	(55.0)	58.9	(56.5)	59.2	2.7
#84	29	(75.0)	73.6	24	(70.0)	73.6	(72.5)	73.6	1.1
#85	35	(83.0)	74.6	24	(71.0)	74.6	(77.0)	74.6	-2.4
#86	31	(34.0)	36.0	42	(37.0)	36.0	(35.5)	36.0	0.5
#87	7	(67.0)	65.9	43	(65.0)	65.9	(66.0)	65.9	-0.1
#88	32	(67.0)	65.1	25	(60.0)	59.4	(63.5)	62.2	-1.3
#89	32	(65.0)	63.1	25	(60.0)	59.4	(62.5)	61.2	-1.3
#90	32	(74.0)	71.9	25	(72.0)	71.1	(73.0)	71.5	-1.5
#91	32	(50.0)	48.4	25	(53.0)	52.5	(51.5)	50.5	-1.0

#92	32	(40.0)	38.6	25	(42.0)	41.8	(41.0)	40.2	-0.8
#93	9	(54.0)	56.8	26	(57.0)	56.9	(55.5)	56.8	1.3
#94	9	(53.0)	55.8	26	(62.0)	61.8	(57.5)	58.8	1.3
#95	30	(72.0)	70.8	26	(70.0)	69.7	(71.0)	70.2	-0.8
#96	30	(70.0)	68.9	26	(53.0)	53.0	(61.5)	60.9	-0.6
#97	34	(60.0)	52.2	26	(54.0)	54.0	(57.0)	53.1	-3.9
#98	6	(48.0)	52.0	28	(50.0)	48.4	(49.0)	50.2	1.2
#99	12	(63.0)	63.4	28	(60.0)	58.2	(61.5)	60.8	-0.7
#100	13	(64.0)	62.2	28	(62.0)	60.2	(63.0)	61.2	-1.8
#101	30	(45.0)	44.4	28	(52.0)	50.4	(48.5)	47.4	-1.1
#102	34	(72.0)	64.0	28	(62.0)	60.2	(67.0)	62.1	-4.9
#103	11	(0.0)	0.2	29	(0.0)	0.2	(0.0)	0.2	0.2
#104	3	(63.0)	55.7	44	(40.0)	47.7	(51.5)	51.7	0.2
#105	3	(72.0)	64.5	44	(65.0)	72.1	(68.5)	68.3	-0.2
#106	3	(62.0)	54.8	44	(50.0)	57.5	(56.0)	56.1	0.1
#107	3	(57.0)	49.9	44	(40.0)	47.7	(48.5)	48.8	0.3
#108	9	(53.0)	55.8	30	(58.0)	57.1	(55.5)	56.4	0.9
#109	18	(75.0)	73.7	30	(71.0)	69.8	(73.0)	71.8	-1.2
#110	26	(78.0)	77.5	30	(73.0)	71.8	(75.5)	74.6	-0.9
#111	26	(65.0)	64.8	30	(66.0)	64.9	(65.5)	64.9	-0.6
#112	17	(45.0)	48.6	45	(46.0)	48.1	(45.5)	48.3	2.8
#113	25	(54.0)	53.5	45	(52.0)	54.0	(53.0)	53.7	0.7
#114	4	(74.0)	72.3	32	(73.0)	70.9	(73.5)	71.6	-1.9
#115	4	(70.0)	68.4	32	(70.0)	68.0	(70.0)	68.2	-1.8
#116	25	(58.0)	57.4	32	(58.0)	56.2	(58.0)	56.8	-1.2
#117	9	(59.0)	61.6	34	(70.0)	62.0	(64.5)	61.8	-2.7
#118	13	(51.0)	49.4	34	(64.0)	56.1	(57.5)	52.8	-4.7
#119	20	(45.0)	44.9	34	(40.0)	32.7	(42.5)	38.8	-3.7
#120	30	(57.0)	56.1	34	(72.0)	64.0	(64.5)	60.1	-4.4
#121	27	(80.0)	79.9	46	(75.0)	79.9	(77.5)	79.9	2.4

TABLE 6
SUMMARY DATA SORTED BY RATER, FOR THE SAME DATASET SHOWN IN TABLE 5 (COMPLETE DATA OF CHAN [11]).

Rater	<--Means-->		Δ	<-Std Dev->		Δ	Total	Paired Raters
1	(66.0)	67.3	1.3	(8.0)	7.8	-0.2	2	2
2	(52.3)	53.7	1.5	(30.3)	29.7	-0.6	4	4
3	(63.5)	56.2	-7.3	(5.4)	5.3	-0.1	4	4
4	(71.8)	70.1	-1.7	(2.9)	2.8	-0.1	4	4
5	(67.0)	70.1	3.1	(6.5)	6.4	-0.1	10	10
6	(61.5)	65.2	3.7	(9.9)	9.7	-0.2	11	11
7	(67.0)	65.9	-1.1	(0.0)	0.0	0.0	1	1
8	(65.0)	59.3	-5.7	(0.0)	0.0	0.0	1	1
9	(61.5)	64.1	2.6	(8.3)	8.1	-0.2	13	13
10	(59.0)	56.1	-2.9	(12.7)	12.4	-0.3	7	7
11	(45.0)	44.3	-0.7	(31.9)	31.2	-0.7	3	3
12	(57.8)	58.3	0.5	(21.6)	21.2	-0.5	9	9
13	(62.5)	60.7	-1.8	(10.4)	10.1	-0.2	16	16
14	(66.0)	65.2	-0.8	(6.0)	5.9	-0.1	2	2
15	(70.0)	63.8	-6.2	(0.0)	0.0	0.0	1	1
16	(43.0)	36.1	-6.9	(0.0)	0.0	0.0	1	1
17	(50.1)	53.6	3.5	(21.0)	20.5	-0.4	8	8
18	(63.0)	62.0	-1.0	(11.7)	11.4	-0.2	8	8
19	(50.5)	51.2	0.7	(15.5)	15.2	-0.3	2	2
20	(63.5)	63.1	-0.5	(8.0)	7.8	-0.2	11	11
21	(61.7)	65.6	3.9	(15.9)	15.6	-0.3	10	10
22	(65.8)	64.7	-1.1	(17.3)	16.9	-0.4	11	11
23	(60.1)	62.2	2.1	(6.4)	6.3	-0.1	8	8
24	(68.0)	71.7	3.7	(6.5)	6.4	-0.1	5	5
25	(57.4)	56.8	-0.6	(6.8)	6.7	-0.1	11	11

26	(65.6)	65.4	-0.3	(8.6)	8.5	-0.2	16	16
27	(68.7)	68.8	0.2	(9.8)	9.6	-0.2	3	3
28	(63.5)	61.7	-1.8	(10.3)	10.1	-0.2	10	10
29	(37.5)	36.9	-0.6	(37.5)	36.7	-0.8	2	2
30	(57.4)	56.5	-0.9	(19.3)	18.9	-0.4	12	12
31	(34.0)	36.0	2.0	(0.0)	0.0	0.0	1	1
32	(62.1)	60.3	-1.8	(11.2)	11.0	-0.2	8	8
33	(78.0)	78.5	0.5	(0.0)	0.0	0.0	1	1
34	(65.8)	57.9	-7.9	(11.2)	11.0	-0.2	8	8
35	(56.5)	48.7	-7.8	(26.5)	25.9	-0.6	2	2
36	(66.5)	77.1	10.6	(1.5)	1.5	-0.0	2	2
37	(73.0)	68.7	-4.3	(0.0)	0.0	0.0	1	1
38	(35.0)	22.7	-12.3	(0.0)	0.0	0.0	1	1
39	(63.0)	56.4	-6.6	(0.0)	0.0	0.0	1	1
40	(70.0)	70.1	0.1	(0.0)	0.0	0.0	1	1
41	(60.0)	66.2	6.2	(0.0)	0.0	0.0	1	1
42	(37.0)	36.0	-1.0	(0.0)	0.0	0.0	1	1
43	(65.0)	65.9	0.9	(0.0)	0.0	0.0	1	1
44	(48.8)	56.2	7.5	(10.2)	10.0	-0.2	4	4
45	(49.0)	51.0	2.0	(3.0)	2.9	-0.1	2	2
46	(75.0)	79.9	4.9	(0.0)	0.0	0.0	1	1

TABLE 7
OUTPUT FROM THE PROGRAM USING DATA OF CHAN [11], SUBSET #3 ONLY.

Project	Rater#1	<--Marks-->		Rater#2	<--Marks-->		<-Averages>		Δ
#7	13	(58.0)	53.6	5	(60.0)	62.9	(59.0)	58.3	-0.7
#8	20	(80.0)	82.4	5	(76.0)	78.5	(78.0)	80.5	2.5
#9	21	(75.0)	77.7	5	(70.0)	72.7	(72.5)	75.2	2.7
#10	22	(78.0)	74.8	5	(75.0)	77.5	(76.5)	76.1	-0.4
#11	22	(73.0)	69.8	5	(68.0)	70.7	(70.5)	70.3	-0.2
#12	26	(77.0)	68.6	5	(76.0)	78.5	(76.5)	73.5	-3.0
#13	28	(74.0)	66.6	5	(60.0)	62.9	(67.0)	64.8	-2.2
#15	13	(60.0)	56.3	6	(55.0)	50.9	(57.5)	53.6	-3.9
#16	21	(83.0)	81.9	6	(73.0)	84.4	(78.0)	83.1	5.1
#17	21	(77.0)	78.7	6	(72.0)	82.5	(74.5)	80.6	6.1
#18	22	(80.0)	76.7	6	(65.0)	69.5	(72.5)	73.1	0.6
#19	22	(83.0)	79.7	6	(72.0)	82.5	(77.5)	81.1	3.6
#20	28	(70.0)	64.2	6	(60.0)	60.2	(65.0)	62.2	-2.8
#21	28	(79.0)	69.7	6	(65.0)	69.5	(72.0)	69.6	-2.4
#22	34	(68.0)	59.6	6	(66.0)	71.4	(67.0)	65.5	-1.5
#23	18	(67.0)	68.8	9	(68.0)	66.4	(67.5)	67.6	0.1
#24	20	(60.0)	59.8	9	(65.0)	65.4	(62.5)	62.6	0.1
#25	22	(70.0)	66.9	9	(59.0)	63.5	(64.5)	65.2	0.7
#26	26	(71.0)	66.4	9	(72.0)	67.7	(71.5)	67.0	-4.5
#27	26	(75.0)	67.8	9	(71.0)	67.4	(73.0)	67.6	-5.4
#28	26	(73.0)	67.1	9	(69.0)	66.7	(71.0)	66.9	-4.1
#29	30	(55.0)	59.2	9	(54.0)	61.9	(54.5)	60.5	6.0
#31	13	(65.0)	62.9	10	(68.0)	64.8	(66.5)	63.9	-2.6
#32	13	(30.0)	16.4	10	(30.0)	11.3	(30.0)	13.9	-16.1
#33	22	(55.0)	52.1	10	(60.0)	53.5	(57.5)	52.8	-4.7
#34	26	(61.0)	62.7	10	(62.0)	56.3	(61.5)	59.5	-2.0
#35	34	(80.0)	67.9	10	(70.0)	67.6	(75.0)	67.7	-7.3
#38	20	(62.0)	62.1	12	(52.0)	61.3	(57.0)	61.7	4.7
#39	21	(62.0)	70.9	12	(58.0)	63.5	(60.0)	67.2	7.2
#40	21	(72.0)	76.1	12	(73.0)	69.2	(72.5)	72.6	0.1
#41	26	(73.0)	67.1	12	(68.0)	67.3	(70.5)	67.2	-3.3
#43	6	(60.0)	60.2	13	(60.0)	56.3	(60.0)	58.3	-1.7
#44	12	(72.0)	68.8	13	(68.0)	66.9	(70.0)	67.8	-2.2
#45	22	(70.0)	66.9	13	(69.0)	68.2	(69.5)	67.6	-1.9

#46	22	(80.0)	76.7	13	(70.0)	69.6	(75.0) 73.1	-1.9
#47	26	(70.0)	66.0	13	(58.0)	53.6	(64.0) 59.8	-4.2
#48	28	(50.0)	52.0	13	(65.0)	62.9	(57.5) 57.5	-0.0
#55	13	(77.0)	78.9	18	(68.0)	69.3	(72.5) 74.1	1.6
#56	21	(50.0)	64.6	18	(59.0)	64.6	(54.5) 64.6	10.1
#57	21	(30.0)	54.1	18	(35.0)	52.1	(32.5) 53.1	20.6
#58	23	(65.0)	66.1	18	(65.0)	67.7	(65.0) 66.9	1.9
#59	26	(51.0)	59.0	18	(62.0)	66.2	(56.5) 62.6	6.1
#62	5	(60.0)	62.9	20	(68.0)	68.9	(64.0) 65.9	1.9
#63	9	(74.0)	68.3	20	(68.0)	68.9	(71.0) 68.6	-2.4
#64	10	(67.0)	63.4	20	(60.0)	59.8	(63.5) 61.6	-1.9
#65	12	(74.0)	69.5	20	(65.0)	65.5	(69.5) 67.5	-2.0
#66	12	(60.0)	64.3	20	(60.0)	59.8	(60.0) 62.1	2.1
#67	13	(66.0)	64.3	20	(66.0)	66.6	(66.0) 65.4	-0.6
#68	23	(58.0)	63.3	20	(65.0)	65.5	(61.5) 64.4	2.9
#69	13	(72.0)	72.2	21	(45.0)	62.0	(58.5) 67.1	8.6
#70	13	(67.0)	65.6	21	(53.0)	66.2	(60.0) 65.9	5.9
#71	18	(73.0)	71.9	21	(70.0)	75.1	(71.5) 73.5	2.0
#72	6	(40.0)	23.0	22	(20.0)	17.7	(30.0) 20.4	-9.6
#73	10	(56.0)	47.9	22	(55.0)	52.1	(55.5) 50.0	-5.5
#74	30	(69.0)	65.2	22	(60.0)	57.1	(64.5) 61.1	-3.4
#75	5	(63.0)	65.8	23	(66.0)	66.5	(64.5) 66.2	1.7
#76	5	(62.0)	64.9	23	(52.0)	61.0	(57.0) 62.9	5.9
#77	9	(49.0)	60.3	23	(60.0)	64.1	(54.5) 62.2	7.7
#78	26	(60.0)	62.3	23	(60.0)	64.1	(60.0) 63.2	3.2
#79	28	(76.0)	67.9	23	(70.0)	68.1	(73.0) 68.0	-5.0
#80	30	(53.0)	58.3	23	(50.0)	60.2	(51.5) 59.3	7.8
#93	9	(54.0)	61.9	26	(57.0)	61.2	(55.5) 61.5	6.0
#94	9	(53.0)	61.6	26	(62.0)	63.0	(57.5) 62.3	4.8
#95	30	(72.0)	66.4	26	(70.0)	66.0	(71.0) 66.2	-4.8
#96	30	(70.0)	65.6	26	(53.0)	59.7	(61.5) 62.7	1.2
#97	34	(60.0)	54.1	26	(54.0)	60.1	(57.0) 57.1	0.1
#98	6	(48.0)	37.9	28	(50.0)	52.0	(49.0) 44.9	-4.1
#99	12	(63.0)	65.4	28	(60.0)	58.1	(61.5) 61.7	0.2
#100	13	(64.0)	61.6	28	(62.0)	59.3	(63.0) 60.5	-2.5
#101	30	(45.0)	54.9	28	(52.0)	53.2	(48.5) 54.1	5.6
#102	34	(72.0)	62.3	28	(62.0)	59.3	(67.0) 60.8	-6.2
#108	9	(53.0)	61.6	30	(58.0)	60.5	(55.5) 61.0	5.5
#109	18	(75.0)	72.9	30	(71.0)	66.0	(73.0) 69.5	-3.5
#110	26	(78.0)	68.9	30	(73.0)	66.9	(75.5) 67.9	-7.6
#111	26	(65.0)	64.2	30	(66.0)	63.9	(65.5) 64.0	-1.5
#117	9	(59.0)	63.5	34	(70.0)	61.0	(64.5) 62.2	-2.3
#118	13	(51.0)	44.3	34	(64.0)	56.8	(57.5) 50.6	-6.9
#119	20	(45.0)	42.9	34	(40.0)	40.2	(42.5) 41.6	-0.9
#120	30	(57.0)	60.0	34	(72.0)	62.3	(64.5) 61.2	-3.3

TABLE 8
SUMMARY DATA SORTED BY RATER, FOR THE SAME DATA SUBSET SHOWN IN TABLE 7 (DATA OF CHAN [11], SUBSET #3 ONLY).

Rater	<--Means-->	Δ	<-Std Dev->	Δ	Total	Paired Raters
5	(67.0) 69.7	2.7	(6.5) 6.3	-0.2	10	10
6	(61.5) 62.9	1.5	(9.9) 18.5	8.6	11	11
9	(61.5) 64.3	2.8	(8.3) 2.7	-5.6	13	13
10	(59.0) 52.1	-6.9	(12.7) 17.9	5.2	7	7
12	(65.0) 66.1	1.1	(7.5) 2.8	-4.7	8	8
13	(62.5) 59.6	-2.9	(10.4) 13.8	3.4	16	16
18	(63.0) 66.7	3.7	(11.7) 6.1	-5.6	8	8
20	(63.5) 63.8	0.3	(8.0) 9.0	1.0	11	11
21	(61.7) 70.7	9.0	(15.9) 8.3	-7.6	10	10
22	(65.8) 62.8	-3.0	(17.3) 17.0	-0.3	11	11

23	(60.1)	64.2	4.1	(6.4)	2.5	-3.9	8	8
26	(65.6)	64.4	-1.2	(8.6)	3.2	-5.5	16	16
28	(63.5)	60.2	-3.3	(10.3)	6.3	-4.0	10	10
30	(62.6)	62.4	-0.2	(9.0)	3.8	-5.2	11	11
34	(65.8)	58.0	-7.7	(11.2)	7.7	-3.5	8	8