The Computer as a Silent Partner in Essay Scoring

David Navon
Haifa University

Yoav Cohen
NITE, Jerusalem

Abstract

Psychometric measurement based on subjective judgments of performance quality (e.g., essay ratings) is, typically, not very reliable. The subjective judgments are often integrated into a single score by means of the following scoring model: Initially, two independent judgments are conducted; then, if the absolute difference between them is not too large, their mean is used as the score. Otherwise, an additional judgment is conducted, and the score is determined by mean of the third judgment and whichever of the original two is closest to it. Whenever the two judgments are sampled from the same distribution, their mean is an unbiased estimate of the true score. However, quite surprisingly, substituting any one of the judgments according to the scoring model described above would result in increased error variance.

In some domains, such as the rating of short essays, it is possible to attain a high level of agreement between a human judgment and a mechanical judgment (Automatic Essay Scoring – AES) based on fairly simple considerations. Though it is not common practice to rely absolutely on AES, the aforesaid high level of agreement suggests that a model employing the difference between a mechanically generated score and a score generated by a human judge is worth considering.

Accordingly, we propose that the following model be put into practice: In the initial phase, two judgments are obtained, one human and the other mechanical. It follows from the logic described above that a large difference between the two scores indicates the likelihood that the human-generated score is fairly far from the true score. Since, in some situations, the validity of correcting for this by averaging that score with the mechanically-generated one is disputable, the recruiting of another human judge is called for. The overall cost of judgment would be substantially reduced by reducing the considerable rate of scores generated by human judges that have to be corrected in this manner. The study explores the benefits of using this model. The current study, which is based on simulated essays and scores, explores the error of measurement associated with various scoring rules.

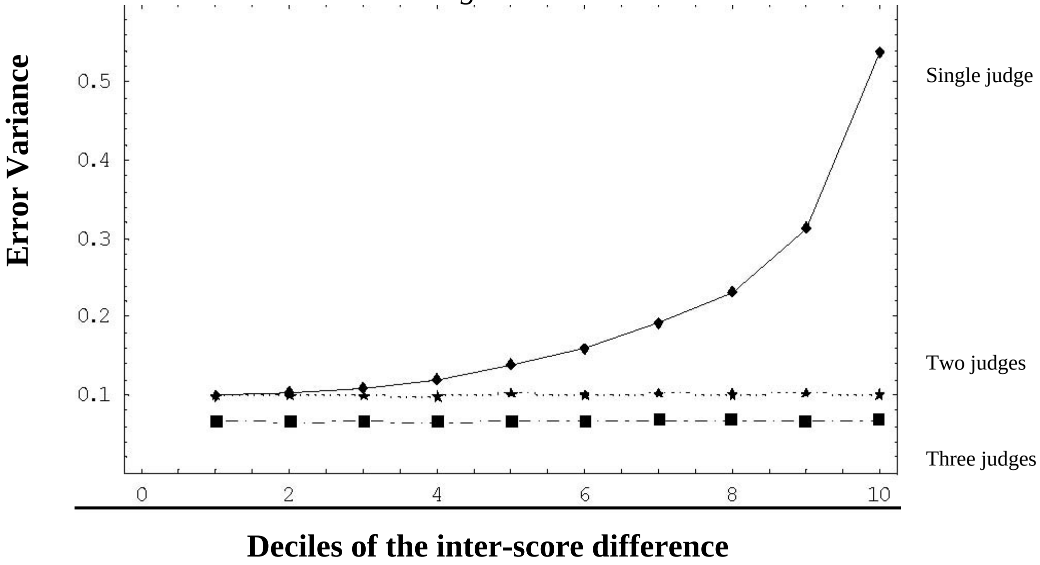
Introduction

Computer-implemented mechanical judgment of solutions to open-ended problems (e.g., essays) is typically considered inadequate and hence inadmissible. On the other hand, should mechanical judgment be highly correlated with human judgment, it might be exploited in a complementary capacity as a means of reducing the cost of human judgments.

Psychometric measurement based on subjective judgments of performance quality (e.g., essay ratings) is, typically, not very reliable. The subjective judgments are often integrated into a single score by means of the following scoring model: Initially, two independent judgments are conducted; then, if the absolute difference between them is not too large, their mean is used as the score. Otherwise, an additional judgment is conducted, and the score is determined by mean of the third judgment and whichever of the original two is closest to it.

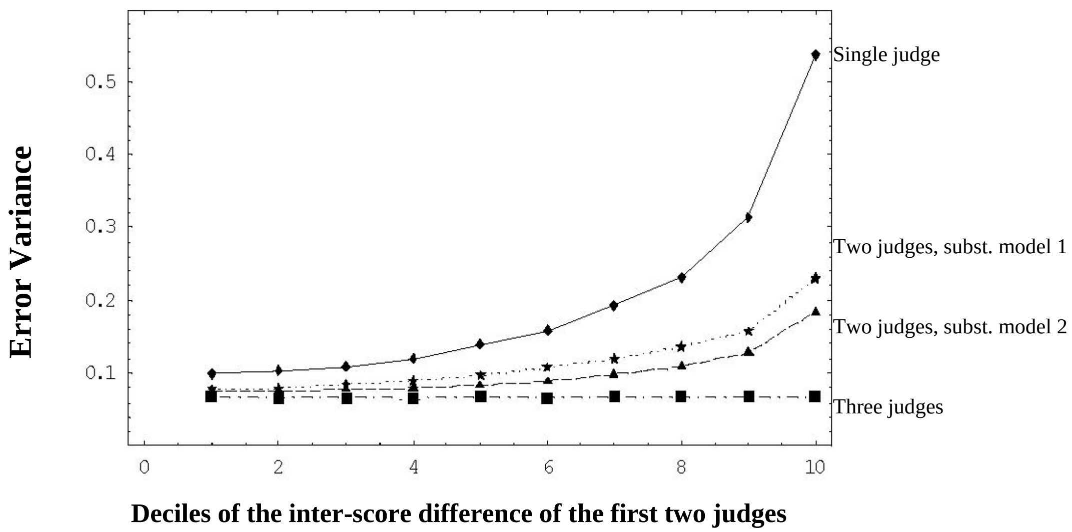
The origins of this model are not entirely clear. Still, its effect on error variance bears examining. Supposing that both original judgments are sampled from a given distribution of scores, each composed of the true score and a judgment error, a large difference between them is typically due to errors that differ in sign and have a high mean of absolute values. If so, there is no necessary correlation, let alone a positive one, between the difference of the errors and their sum. Consider for example a normal distribution of errors. Figure 1 presents the error variance as a function of (deciles of) the inter-score difference of the first two judges. The three curves correspond to three scoring rules – by the first judge only, by the mean of the first two judges and by the mean of three judges. The error variance of the judgment mean is not related to the magnitude of the difference (see the asterisk curve).

Figure 1



Thus, whenever the two judgments are sampled from the same distribution, their mean is an unbiased estimate of the true score, and there is no basis upon which to estimate it a-priori as more distant from the true score than any other mean of two judgments. In any event, substituting any one of the judgments according to the scoring model described above would result in increased error variance. Figure 2 presents the error variance as a function of the inter-score difference between the first two judges for four scoring rules: the first judge only (see the diamonds curve), the mean of the first three judges (see the squares curve) and by two substitution models – the scoring model described above (model 1, see the asterisk curve), and another one in which the additional judgment is averaged with one of the original two, picked at random (model 2, see the triangle curve). Ironically, in the former model, the higher the difference, the larger the increase would be. The latter is mildly superior in that respect.

Figure 2



Thus far we have discussed the shortcomings of the model described above, as well as those of any other substitution model. These shortcomings may easily be avoided by adopting a simpler model, which determines the score by the mean of N judgments, irrespective of differences between them, where N is the smallest number required for attaining a tolerable error variance given cost constraints (see the two lower functions in Figure 1). On the other hand, costs can be considerably spared by using another model, one that employs mechanical judgment and takes differences in judgment scores into account. Such a model is described below.

In some domains, such as the rating of short essays, it is possible to attain a high level of agreement between a human judgment and a mechanical judgment (Automatic Essay Scoring – AES) based on fairly simple considerations. Though it is not common practice to rely absolutely on AES, the aforesaid high level of agreement suggests that a model employing the difference between a mechanically generated score and a score generated by a human judge is worth considering.

Accordingly, we propose that the following model be put into practice: In the initial phase, two judgments are obtained, one human and the other mechanical. It follows from the logic described above that a large difference between the two scores indicates the likelihood that the human-generated score is fairly far from the true score. Since, in some situations, the validity of correcting for this by averaging that score with the mechanically-generated one is disputable, the recruiting of another human judge is called for. The overall cost of judgment would be substantially reduced by reducing the considerable rate of scores generated by human judges that have to be corrected in this manner. The study explores the benefits of using this model. A similar study was conducted by Bridgeman (2005) on actual ratings of GRE essays. The current study, which is based on simulated essays and scores, explores the error of measurement associated with various scoring rules.

Method

This was a simulation study. Simulated ratings to simulated essays were generated by using the standard add-on packages for statistics of the *Mathematica* system (Wolfram Research, 1999). Among other things, the *Mathematica* system is known for its accuracy in analyzing large amounts of data consisting of high precision observations.

Each simulation run, which typically was based on ratings of 100,000 simulated essays, consisted of the following stages:

1. Generating a set of true scores for the essays. The true scores were normally distributed.
2. For each true score a triad of observed scores was generated by adding an error component (independent of the true score) to the true score. The first two scores were the two original ratings of the essay, one by a human rater and the other by an AES system. The third score represented the additional human rating which could, in principle, either replace, or be averaged with, the first human score.
3. The absolute difference between the first two scores was calculated and the score triads were ordered by this value, from smaller to larger difference between scores.
4. The addition of a third score (second human score) was applied according to the specific research question.

As is evident from the description of the simulation procedure, the data conform to classical test theory – the observed score is the sum of two components: a true score and an uncorrelated error component.

The following factors were manipulated in order to study the effects of various procedures on the error of measurement:

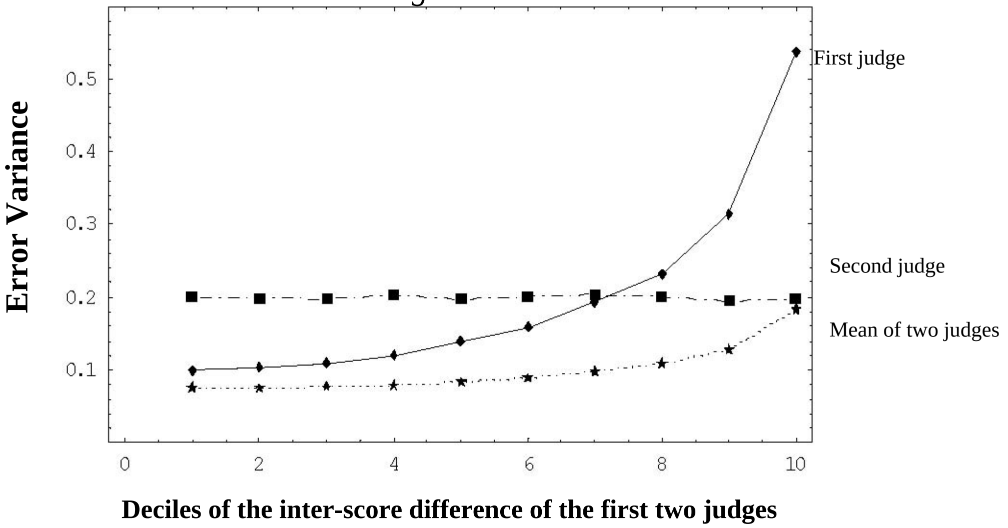
1. The absolute difference between any two scores.
2. The rule used for scoring the essays.

Results

Each simulation was performed on 100,000 essays, for which true scores were normally distributed with a zero mean. The errors were also normally distributed with zero mean and variance of 0.20.

Figure 3 presents three curves, each of which plots the error variance resulting from each of three scoring rules – by the first judge only (see the diamonds curve), by the second judge only (filled squares), or by the mean of both (asterisks), as a function of the difference between the score of the first judge and the mechanically-generated score (grouped into 10 deciles). One obvious result emerging from the figure is that when the differences are small, the reduction in error effected by using the score generated by the second judge as well is very slight. Furthermore, it seems fairly evident that with especially small differences, using the first score alone is not much worse than using the mean of the two. Hence, it seems eminently justifiable to use the score of the first judge only in some cases (see the following section). In other cases, an additional judge has to be recruited. His or her score would have to be averaged with the score yielded by the first judge.

Figure 3



Another option is weighting (not necessarily equally) the two human-generated scores. Figure 4 presents five different weightings – in which the additional score is weighted 0, .25, .50, .75, or 1.0 as a function of the difference between the score of the first judge and the mechanically-generated score. As can be seen in the figure, a .75 weighting (plotted here in the squares curve) is superior to simple averaging (plotted here in the asterisk curve) in about ten percent of the cases.

Figure 4

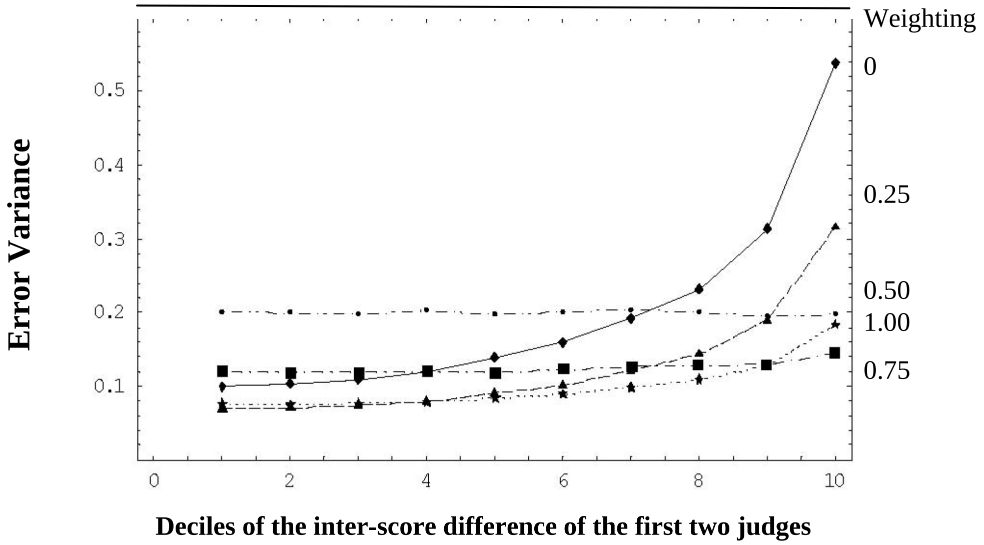
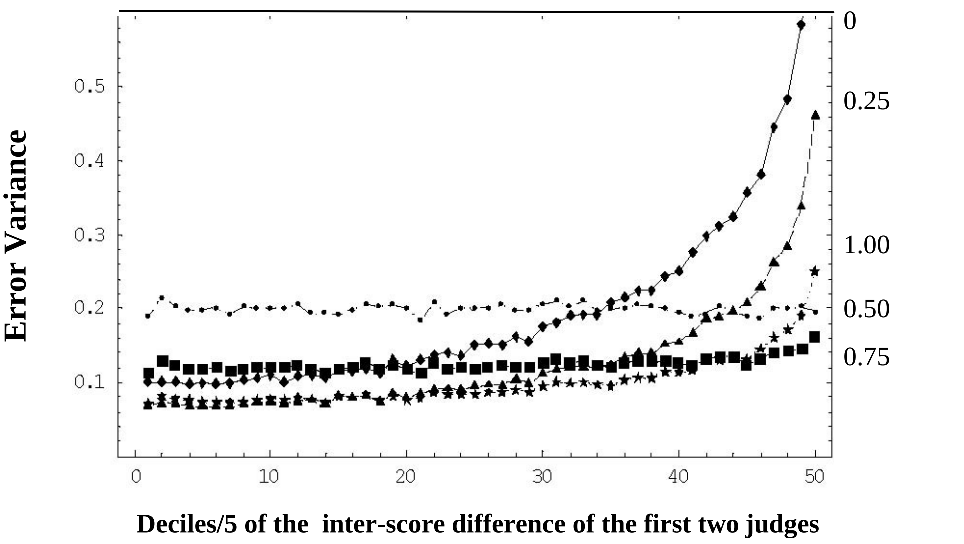


Figure 5 presents the same curves that appear in Figure 4, at a higher resolution (each data point is based on 2% of the cases). It shows that when the difference is particularly high, in about 2% of the cases, the score generated by the additional judge (plotted here in the circle curve) is itself superior to simple averaging – not only when it is weighted by .75. The reason is that in such a case it is highly likely that the score generated by the first judge is so far from the true score that even the average of it and the additional score is farther from the true score than the expected additional score itself.

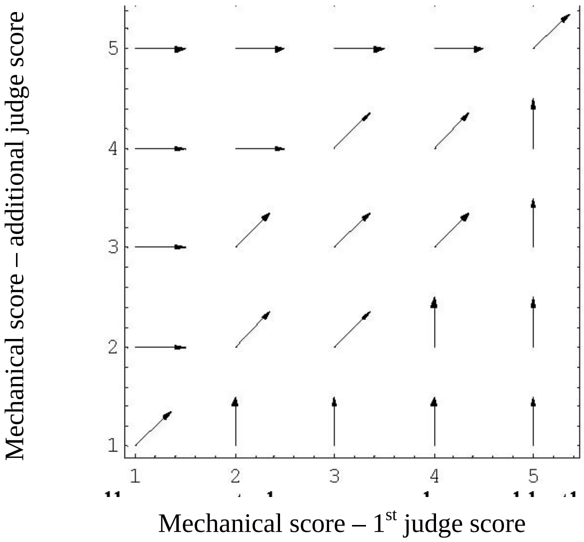
The asymmetry between the scores that is evident in Figures 4 and 5 might at first seem puzzling. It should, however, be borne in mind that it does not reflect a true asymmetry, rather one brought about by the exposition in these figures. The position on the abscissa is directly related to the extent of the difference between the score of the first judge and the mechanically-generated score. When that difference is particularly high, it is likely that the former has almost no diagnostic value. Since there is no a-priori difference between the first judge and the additional one, there are of course cases in which the difference between the score of the additional judge and the mechanically-generated score is especially high, so high as to render the former devoid of any practical value. However, these cases are scattered throughout the horizontal extent of the diagram.

Figure 5



When an additional third judge is recruited, optimal weighting of the two scores depends on the differences between the mechanically-generated score and the two human-generated ones. Figure 6 presents regions of optimal weighting (.25, .50 or .75 for the additional score, marked by horizontal, diagonal or vertical arrow respectively) within a plane spanned by (a) the difference between the mechanically-generated score and the score of the first judge, (b) the difference between the former score and the score of the additional judge.

Figure 6



In sum, mechanically-generated scores can be used both for reducing the mean number of human judges and for judicious weighting of scores generated by two human judges.

Conclusion

Parents and educators do not always approve of using AES systems for educational purposes, let alone in high stakes testing programs. It is, thus, fortunate that AES can be also used to benefit reliability of scoring by human raters only. It can be used for flagging cases in which an additional rating of the essay is required, and it can be used to adjust the relative weight of two ratings or scores.

References

Bridgeman, B (2004) *E-rater as a quality control on human scorers.* Presentation in the ETS research Colloquium series, Princeton, NJ. Cited in ETS R&D Connections, April, 2005.

Wolfram Research, Inc. (1999) *Mathematica Version 4.0.* Champaign, IL.