Generating Rubric Scores From Pairwise Comparisons

Shayne Miel, David Adamson, Bronwyn Woods - Turnitin

Holly Garner - EverEd

Abstract

Using pairwise comparisons, or comparative judgement scoring, to score essays on a holistic rubric is potentially a more reliable and less expensive scoring method than traditional handscoring. However, a challenge it presents is how to generate discrete rubric-level scores from the continuous output of the ranking process. We establish metrics for measuring the validity, reliability, and consistency of the comparative judgement scoring process, and use them to compare methods for assigning discrete rubric scores to the resulting ranked list.

When assessing constructed response items, the standard practice is to have a set of expert readers score the essays based on a rubric. However, this method, even with the common model of two independent reads and a resolution score, has trouble achieving high levels of scoring reliability (Brown, 2004). The reliability of this handscoring method is hampered by several factors. First, there is the cognitive load of assigning a numeric score to a trait of a student's writing. This effect is especially pronounced for essays that lie on the border between score points. Second, when readers know that they are being judged on their ability to agree with one another, they tend to score towards the middle of the range, shying away from applying the extreme scores. In addition to these issues, there is a large upfront cost in rangefinding, developing anchor sets, and training the readers when doing traditional handscoring. Generally, this cost is distributed across tens or hundreds of thousands of essays, but the upfront cost makes scoring small data sets expensive. This is especially true when developing training sets for automated scoring, where only a few hundred scores are needed.

Comparative judgement scoring offers a promising alternative to traditional rubric handscoring by offering a simpler task to the readers, which in turn can lead to more reliable and less expensive scoring. Generating a ranked list from pairwise comparisons is a common psychometric practice (Thurstone, 1927), that has been used in constructed response scoring (Heldsinger, 2010). Instead of presenting the reader with a single essay and asking for a rubric score, the reader is shown two essays and simply asked which one is better according to the rubric. These comparisons are then collected and used to rank the essays along a

continuous scale. There are multiple techniques for turning pairwise comparisons into continuous scores. For this paper, we use the SVM Ranking method described by Wauthier, Jordan, and Jojic (2013). The number of necessary pairwise comparisons is an open question. Ideally, for n essays we would generate $n^2 * d$ comparisons, where d is a redundancy factor that balances out noise in the individual comparisons. However, even collecting n^2 pairwise comparisons is prohibitively expensive. In this study we use $n \log n$ randomly selected comparisons which provides a good approximation of the true rank order (Jamieson, 2011).

One of the challenges for using comparative judgement on a rubric-scored item, however, is determining how to set the cut points in the ranked list of essays to get the whole-number rubric scores. Previous work has used Markov Chain Monte Carlo (MCMC) estimation to fit a Bradley-Terry model with anchor papers and an assumed prior distribution (Steedle, 2014). The continuous ranking scores generated by this method align with the rubric scores given by the anchor papers, and can simply be rounded to get the whole-number rubric scores. However, this approach suffers from several drawbacks. First, it requires that the anchor papers properly cover the range of each score point. That is, the anchor papers must lie as close as possible to the desired cut points, which leads to a chicken/egg scenario when the intent is to discover those cut points from the data. The other problem with this approach is that one must specify the expected number of essays at each score (the prior). From experience we know that constructed response scoring tends to lie along a bell-shaped curve, but that is certainly not the case for every dataset/rubric pair. Ideally we would like to find a method that is independent of prior assumptions about the data and free of the need for developing anchor sets, which can be a time-consuming and error-prone process.

The comparative judgement process used in this paper can be described as follows:

1. Pairwise comparisons between random pairs of essays are collected, indicating which essay should score higher on a given rubric trait.
2. These comparisons are used to train a support vector machine (SVM). The weights of the SVM's separating hyperplane become the continuous score values for the essays.
3. The essays are ranked based on the continuous score values from the previous step.
4. Some method is used to transform the ranked continuous scores into discrete rubric scores.

In the remainder of this paper we describe several methods for transforming the ranked continuous scores into discrete rubric scores (step 4 above). We then compare these methods with one another and traditional handscoring, considering dimensions of validity, reliability, and consistency.

Data

We use essays collected from Turnitin's RevisionAssistant[1], an online writing and feedback tool, in which 6th-8th grade students were asked to respond to two open-ended writing prompts (one informative writing and one persuasive writing). Approximately 500 submitted essays were taken from each prompt and scored by an independent handscoring vendor on a 4-point rubric with 2 traits, using both the traditional and the comparative judgement method. For the traditional method, essays were double-scored with resolution on non-exact agreement. For the comparative judgement method, approximately 3,500 comparisons were collected per trait, with pairings selected randomly such that each essay was compared to exactly 14 others.

Methods

Our focus in this work is comparing several methods for transforming the essay ordering given by the comparative judgement method into discrete rubric scores. We propose a clustering method, and compare it against several simpler alternatives. In all cases, we begin with the ranked (real number) scores generated by the pairwise comparison process. Our task is to find the cut points that allow us to transform the ranked scores into whole-number rubric scores in the range 1-4. The first two naïve methods, **uniform** and **normal**, assume prior knowledge of an appropriate scoring distribution, while the third and fourth, **scale and round** and **cluster**, do not, relying only on the continuous rank scores. The four approaches are described below:

Uniform

We assume that the scores are uniformly distributed across the data set. For a 4-point rubric, cut points are set at 25th, 50th, and 75th percentiles of the ranked list of essays.

Normal

We assume the scores are distributed according to an underlying normal distribution, with mean at the midpoint of the score range and all scores within one standard deviation of the mean. From prior experience with these rubrics, we set $\mu = 2.5$ and $\sigma = 1$. For our 4-point rubric, this means that the cut points are set at 13.4%, 50%, and 86.6% of the ranked list of essays.

Scale and Round

Because the ranked scores exist along a finite number line, we can simply scale them to lie between 0.5 and 4.5, and then round them to get the desired whole-number rubric scores.

[1] www.revisionassistant.com

Cluster

We attempt to discover natural groupings of the real-number ranked scores along the continuous number line. Unsupervised clustering is a machine learning technique that attempts to group data so that points that are near to one another in some space belong to the same cluster. We can use the continuous scores generated by the pairwise comparison process to cluster the essays into as many groups as there are rubric scores, giving us the desired whole-number scores without assuming any prior information about the distribution of the scores. In particular, we use the Jenks Natural Breaks algorithm, which was originally used to color choropleth maps (Jenks, 1977). This is a dynamic programming algorithm that maximizes the variance between groups while minimizing the variance within groups. It can be shown that when using the algorithm on one-dimensional data (as is the case when clustering the rank scores), it is equivalent to the optimal k-means clustering.

Evaluation

We are evaluating several techniques for transforming continuous comparative judgement rankings into discrete rubric scores. As with any new scoring process, it is important to determine whether the methods are both valid and reliable. In this case, we want to know whether the final discrete scores appropriately measure the construct under examination (**validity**) and whether a repetition of the scoring process with different readers would result in the same score for the same essay (**reliability**). Because we are primarily interested in this process as a means of generating training sets for automated scoring, we add a third criteria: **consistency**. Given a standard modeling process, do the essays and scores from a test item and scoring process carry enough signal to use the data operationally in an automated scoring system?

Validity of a test item or scoring process is generally measured by showing per-student agreement with some external measure of the writing construct (scores on a multiple-choice section of the test, for instance). However, this kind of external measure is very difficult to collect. We instead use the fact that the traditional handscoring process is a standard industry practice with many years of validity testing behind it. For each of our experimental methods, we examine the level of agreement between the comparative judgement rubric scores and the traditional handscoring scores.

Reliability is typically measured via inter-rater reliability (IRR), the level of agreement between two independent readers on the essay scores. It is straightforward to determine this under the traditional handscoring method, but it is less clear how to calculate IRR when doing pairwise comparison. Ideally, you would have two sets of readers each make $n \log n$ comparisons and measure the agreement between the rubric scores generated from each set of comparisons. To approximate this, we split our comparison data in half and compare rubric scores generated from each set of $n * (\log n)/2$ comparisons. Because $n * (\log n)/2$ is too few comparisons to generate a good approximation of the ranking, we expect

that the these numbers will be lower than they would be if properly measured with two sets of n log n comparisons.

We determine the consistency of a candidate set of holistic rubric scores on constructed response essay writing empirically. We train a machine learning model on a subset of the generated scores from the candidate method. We then measure the ability of the model to predict the remaining scores. We repeat this process (cross-validation) to estimate ability of the model to learn the set of scores under consideration. We call this cross-validated measure of predictive accuracy "consistency".

Though many modeling approaches are possible, we use the model described by the open-source LightSide machine learning system (Mayfield and Rosé, 2013). While consistency does not tell us anything about how close the candidate set of scores is to the "true" scores, when combined with a validity and reliability measure, a high cross-validated agreement does indicate that there is some signal to the assigned scores. One possible interpretation of this measure is that high consistency shows that similar essays receive similar scores[2].

In all cases where we measure the similarity or accuracy of one set of scores against another, we use quadratic weighted kappa as the primary metric. Because there is randomness involved in the reliability and consistency measures, we repeat the process 20 times for each metric and show the mean and the 99% confidence interval.

Results

Writing Genre	Trait	Uniform	Normal	Scale and Round	Cluster
Informative	Clarity & Focus	0.577	**0.601**	0.597	0.590
	Use of Evidence	0.564	0.628	**0.638**	0.562
Persuasive	Analysis & Organization	0.616	**0.643**	0.597	0.642
	Language & Genre Awareness	0.567	0.578	**0.613**	0.591

Table 1: Quadratic weighted kappa between scores generated by each method and the traditional handscoring process

[2] For some definition of "similar" as modeled by the automated scoring engine.

Writing Genre	Trait	Traditional	Uniform	Normal	Scale and Round	Cluster
Informative	Clarity & Focus	0.579	**0.619 ± 0.014**	0.588 ± 0.016	0.571 ± 0.021	0.598 ± 0.015
	Use of Evidence	0.525	0.572 ± 0.015*	**0.572 ± 0.011**	0.554 ± 0.015	0.565 ± 0.021
Persuasive	Analysis & Organization	0.555	**0.605 ± 0.015**	0.598 ± 0.016	0.583 ± 0.014	0.597 ± 0.023
	Language & Genre Awareness	**0.560**	0.556 ± 0.013	0.545 ± 0.015	0.522 ± 0.023	0.527 ± 0.022

Table 2: Inter-rater reliability for traditional handscoring and approximate inter-rater reliability for the comparative judgement methods (QWK)[3], with 99% confidence intervals

Writing Genre	Trait	Traditional	Uniform	Normal	Scale and Round	Cluster
Informative	Clarity & Focus	0.564 ± 0.011	0.691 ± 0.010	0.669 ± 0.011	0.655 ± 0.013	**0.713 ± 0.007**
	Use of Evidence	0.613 ± 0.011	0.676 ± 0.017	0.705 ± 0.013	0.676 ± 0.019	**0.706 ± 0.009**
Persuasive	Analysis & Organization	0.675 ± 0.006	0.746 ± 0.015	0.730 ± 0.014	0.722 ± 0.010	**0.772 ± 0.008**
	Language & Genre Awareness	0.618 ± 0.011	**0.716 ± 0.017**	0.707 ± 0.015	0.687 ± 0.009	0.714 ± 0.008

Table 3: Quadratic weighted kappa between rubric scores generated by each method and predicted scores from a model trained on those scores, via 10-fold cross-validation, with 99% confidence intervals

Discussion

Our results show that the comparative judgement process is both valid and reliable, and that it leads to more accurate automated scoring models than traditional handscoring. While no particular rubric score generation method outperforms all of the others on validity and reliability, all of the methods perform within an acceptable range. When looking at the consistency metric,

[3] Note that the approximate comparative judgement IRR is lower here than in actuality because of the limited data available to sample the pairwise comparisons.

however, the models trained with the clustering approach are noticably better than the other methods (Table 3). Because the clustering method is free of assumptions about the true distribution of the scores, the consistency that it offers provides a strong argument for its use in practical applications. Future work should examine whether performance remains as high on badly skewed data and non-normal score distributions, as well as whether the consistency measure remains high with other kinds of automated scoring models.

The validity measurements in Table 1 should be taken with a grain of salt. It is a complicated position to claim that a) traditional handscoring is not as accurate or reliable as we would like it to be and b) we should measure the validity of new approaches by comparing them to the traditional scores. However, with one exception, the validity numbers are all higher than the traditional scoring inter-rater reliability numbers (Table 2). This means that these processes agree with the traditional scores at least as well as two expert readers agree with one another. Future work should try to measure validity against an external assessment of the writing construct to verify that scores from the comparative judgement process are, in fact, *more* valid than the traditional scores.

References

Brown, Gavin TL, Kath Glasswell, and Don Harland. "Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system." *Assessing Writing* 9.2 (2004): 105-121.

Heldsinger, Sandra, and Stephen Humphry. "Using the method of pairwise comparison to obtain reliable teacher assessments." *The Australian Educational Researcher* 37.2 (2010): 1-19.

University of Kansas. Dept. of Geography, and G. F. Jenks. *Optimal data classification for choropleth maps*. 1977.

Johnson, Robert L., James Penny, and Belita Gordon. "Score resolution and the interrater reliability of holistic scores in rating essays." *Written Communication* 18.2 (2001): 229-249.

Mayfield, Elijah, and C. P. Rosé. "LightSIDE: open source machine learning for text." In *Handbook of Automated Essay Evaluation,* eds. Shermis, Mark and Burstein, Jill. Routledge (2013).

Steedle, Jeffery, and Ferrara, Steve. "A Comparative Judgment Approach to Essay Scoring". Presented at the California Educational Research Association Annual Conference. San Diego, CA. 4 Dec. 2014. Conference Presentation.

Thurstone, Louis L. "A law of comparative judgment." *Psychological review* 34.4 (1927): 273.

Wauthier, Fabian, Michael Jordan, and Nebojsa Jojic. "Efficient ranking from pairwise comparisons." *Proceedings of the 30th International Conference on Machine Learning*. 2013.