Using Adaptive Comparative Judgement to Assess Student Work in an MBA Course

Matthew Metzgar
University of North Carolina at Charlotte, USA

Abstract

A common instructional problem with large classes is the assessment of non-standardized student work such as short-answer questions, papers, and projects. The large grading load associated with assessing open-ended work often leads to a greater reliance on multiple-choice questions. While multiple-choice questions provide beneficial information about student performance, they may fail to capture other elements of student performance in regards to communication, writing, and generation of ideas.

One potential solution to this problem is the concept of adaptive comparative judgment (ACJ). ACJ is based on the simple premise that while peers may not have the ability to place an objective grade on a paper, they can competently compare two different papers and choose which is superior. With a large enough number of student "judgements" on a body of peer work, the collective results from the comparison process can produce rankings that are on par with how the instructor would rank these papers.

This session will highlight the instructor's use of ACJ with an MBA problem-based class. ACJ was used over a number of consecutive assignments, and it produced a high correlation with the instructor's rankings. Students also expressed satisfaction with the system and may have benefited from seeing other (anonymous) student work. ACJ represents a promising and fairly easy-to-use approach for grading open-ended work in large classes.

1. Introduction

With large classes, many instructors turn to multiple-choice questions for assessment. Multiple-choice questions can moderate grading load while serving as a proxy for student performance. While multiple-choice questions provide beneficial information about student performance, they may fail to capture other important elements of student performance in regards to communication, writing, and generation of ideas [1]. Non-standardized student work such as short-answer questions, papers, and projects may be better gauges of student performance but are time-intensive in their use. Instructors may then be conflicted between the need for quality assessment balanced with time constraints.

One potential solution is to involve a student's peers in the assessment of their work. Peer assessment has a long history of research behind it [2]. The obvious issue with peer assessment is how accurate the peer's mark will be compared to that of the instructor. Improving the accuracy of a peer judgment may involve training, repetition, and guidance. Again, this may involve additional time and other resources from the instructor.

An alternative way to use peer judgements is to aggregate various peer judgements together to produce a more accurate assessment. This is, in effect, "crowd-sourcing" an assessment by having a number of peers all assess one object. It may be possible that this crowd-sourcing will produce a quality judgement given a large enough number of peers.

The assessment technique of Adaptive Comparative Judgement (ACJ) is based upon this idea. ACJ is based on the simple premise that while peers may not have the ability to place an objective grade on a paper, they can competently compare two different papers and choose which is superior. With a large enough number of student "judgements" on a body of peer work, the collective results from the comparison process can produce rankings that are on par with how the instructor would rank these papers. From this rank-ordering, an instructor can then assign marks based upon a grading scale or rubric.

ACJ has been utilized in various educational environments and has produced impressive results [3]. Jones & Alcock [4] demonstrated the use of ACJ for an undergraduate calculus course. Seery, Canty, & Phelan [5] successfully used ACJ for assessing design skills. Newhouse & Cooper [6] used ACJ to reliably assess Italian language skills. Tarricone & Newhouse [7] used ACJ to assess creative performance and capability.

Given the potential of ACJ, the instructor implemented Adaptive Comparative Judgement in an MBA problem-solving class. ACJ was used over a number of consecutive assignments to test its validity against instructor rankings. Students were also surveyed in regards to their acceptance and views on the ACJ system. Overall, ACJ shows potential in terms of greater accuracy and greater student participation in assessment.

2. Experiment

The use of ACJ was piloted in an MBA class of 34 students. The MBA program was located at a large, public university in the southeastern United States. ACJ was used over five consecutive written tasks, where students attempted to address a real-world problem or scenario.

Adaptive Comparative Judgement was administered via the web with the free tool, "Mo More Marking" [8]. The workflow process went as follows: students submitted a group paper via the web with no names or identifying features included. Papers were labeled alphabetically and scanned in as pdf files. Papers were then distributed for "student judgements". A student judgement simply consists of viewing two anonymous papers and then choosing which paper students thought was better. All students were assigned two judgements each to achieve the necessary volume of judgments.

In this experiment, students were simply asked to "Choose which paper you think is better". While such a directive may seem vague, previous research shows grading rubrics do not necessarily improve the accuracy of ACJ judgements [9]. Many real-life decisions are also holistic in nature, such as choosing which candidate to vote for in an election, which restaurant to go to for dinner, etc. While individual details behind such decisions are important, ultimately a person has to rank one option above another.

Students were given one week to make their two judgements. Several safety measures are available to ensure that students take their judgements seriously. One tool is that the time a student takes to make a judgment is recorded. Therefore, if a student is judging two multi-page papers in just a few seconds, the judgement may be suspect. Second, there is an "infit" measure for students as judges. This measures if a student's judgement is generally in line with the rest of the class. A student whose judgements are far out of line with the general class may deserve further scrutiny.

The system also produces a reliability coefficient. Values above 0.8 signify the results are stable and repeatable (according to the system guide). Values below this threshold may bring the results into question. Table 1 shows the basic data across the five tasks.

The other major part of this experiment was comparing the rank order produced by the students via the ACJ tool versus the instructor's rank order of the papers. The general theory of ACJ is that with enough student judgements, the rank order produced by the cumulative student judgements will approach that of the expert. In this case, the instructor assessed and rank ordered all student submissions before viewing the ACJ results.

Table 1. Descriptive Data for ACJ Tasks

	Candidates	Judgements	Reliability
Task #1	13	64	0.77
Task #2	12	68	0.81
Task #3	12	65	0.56
Task #4	12	62	0.76
Task #5	8	64	0.82

3. Results

For each task, the rank order of the ACJ was compared to the instructor rank order. Overall, the rankings were surprisingly similar. One area of high agreement was the sorting of top performers and low performers.

In four of the five tasks, the top three and bottom three were identical for the ACJ rankings and the instructor rankings. In one task, a candidate that the instructor had mid-range was ranked in the top 3 by the students. In all five tasks, the bottom three candidates were exactly the same for ACJ and the instructor.

There was some level of disagreement within candidates in the mid-range of the rankings. However, in an absolute sense the differences were quite small. The largest difference of ranking was only two places comparing ACJ and the instructor.

A correlation analysis was performed to compare ACJ ranking versus instructor ranking across each task. The results are presented in Table 2.

Table 2. Correlation between Instructor and ACJ Rankings

	Correlation Coefficient
Task #1	0.74
Task #2	0.77
Task #3	0.55
Task #4	0.79
Task #5	0.83

The correlation generally became tighter as the tasks proceeded. The most obvious explanation is that students became more comfortable and adept as judges. Task #3 appears to be an outlier, as the students felt the assignment instructions were not fully clear. Reliability on Task #3 was also low.

4. Grading

Though this ACJ tool provides a rank order of submissions, the system does not produce a numeric or letter grade. This responsibility still lies with the instructor. This is, of course, a subjective process though the ACJ does provide helpful information in this aspect.

This ACJ tool does provide both the raw information about the number of comparisons "won" by a candidate, and also a "true score" which is a calculated score based on these comparisons. When viewing these true scores, one can see natural breaks in the rankings. For example, there may be three papers clustered together in terms of score, and then a sizeable drop down for the next ranked submission. These natural breaks were used by the instructor to create grade "buckets", which then corresponded to numerical scores (ex. 80, 85, 90, etc.). This is admitted a subjective process; however, this additional data is normally not present when grading papers. Hence, if anything, this additional data may make the grading less subjective than regular grading of papers.

In wanting to give weight to the students work to rank the submissions, the ultimate grade given to a paper was the average of the instructor numerical grade and the ACJ numerical score turned into a numerical grade. For example, the instructor may have initially scored a paper as 80/100. When the ACJ rankings were turned into numerical grades, the ACJ score may have been 85/100. The ultimate grade the student received was the average 82.5/100.

Student feedback to the tool was generally positive. There were no student complaints about the system either verbally or as part of the course evaluations. Students seemed to enjoy having a say in the assessment process. The use of ACJ made assessment a more democratic practice in this case.

5. Conclusion

This paper reviewed the use of an innovative rank ordering system for assessing student work. Adaptive Comparative Judgement (ACJ) combines a volume of student judgements to rank order student submissions. This rank order correlates highly to an expert ordering of the submission.

ACJ was successfully used in an MBA class to make assessment more accurate and democratic. Students expressed satisfaction with the system and may have benefited from seeing other (anonymous) student work. ACJ represents a promising and fairly easy-to-use approach for grading open-ended work in large classes.

6. References

[1] Kuechler, W. L., & Simkin, M. G. (2010). Why Is Performance on Multiple-Choice Tests and Constructed-Response Tests Not More Closely Related? Theory and an Empirical Test. *Decision Sciences Journal of Innovative Education*, *8*(1), 55-73.

[2] Dochy, F. J. R. C., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher education*, *24*(3), 331-350.

[3] Pollitt, A., & Whitehouse, C. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. Center for Education Research and Policy:https://cerp.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_CW_20062012_2.pdf (Access date: 4 June 2016).

[4] Jones, I., & Alcock, L. (2012). Summative peer assessment of undergraduate calculus using adaptive comparative judgement. *Mapping university mathematics assessment practices*, 63-74.

[5] Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, *22*(2), 205-226.

[6] Newhouse, C. P., & Cooper, M. (2013). Computer-based oral exams in Italian language studies. *ReCALL*, *25*(03), 321-339.

[7] Tarricone, P., & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, *13*(1), 1.

[8] Wheadon, C. (2016) 'No More Marking'; www.nomoremarking.com (Access date: 5 June 2016).

[9] Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, *39*(10), 1774-1787.