2023-10-10

Exploring The Reliability, Time Efficiency, And Fairness Of Comparative Judgement In The Admission Of Architecture Students

Lotte VAN DEN HEUVEL
Delft University of Technology, The Netherlands, L.vandenHeuvel@tudelft.nl

Nina Lotte BOHM
Delft University of Technology, The Netherlands, n.l.bohm@tudelft.nl

Recommended Citation

van den Heuvel, L., & Bohm, N. L. (2023). Exploring The Reliability, Time Efficiency, And Fairness Of Comparative Judgement In The Admission Of Architecture Students. European Society for Engineering Education (SEFI). DOI: 10.21427/WEY8-JZ69

EXPLORING THE RELIABILITY, TIME EFFICIENCY AND FAIRNESS OF COMPARATIVE JUDGEMENT IN THE ADMISSION OF ARCHITECTURE STUDENTS

L. van den Heuvel
Delft University of Technology
Delft, The Netherlands

N.L. Bohm
Delft University of Technology
Delft, The Netherlands
ORCID: 0000-0002-5054-9144

***Conference Key Areas**: 11. Innovative Teaching and Learning Methods; 13. Built Environment and Architecture Education*
***Keywords**: Assessment, Comparative judgement, Fairness, Reliability*

ABSTRACT

It is common in architecture education to quantify the quality of assignments into grades, often done by one or two teachers using rubrics. However, this can have several downsides. It suggests an objective preciseness that is debatable for the creative assignments in the field of architecture, and the assessment is dependent on the judgement of only one or two people. Comparative judgement (CJ) offers an alternative to rubric-based assessment by applying pairwise comparison to student assignments, resulting in a ranking instead of a grade.

We used a mixed methods approach to compare the reliability, time efficiency, and fairness of CJ in the selection of students for an undergraduate architecture programme at Delft University of Technology in the Netherlands. Teachers involved in the rubric-based approach for student selection were asked to re-assess a random

selection of the assignments using CJ. Reliability and time investments for both methods were compared, and the involved assessors were asked in a focus group setting which of the two methods they perceived as more reliable and fair. Comparing rubric-based assessment to CJ is new, as previous studies have only looked at these assessment methods in isolation.

Findings indicate that CJ can be serve as a more reliable and time efficient alternative to rubric-based assessment. However, teachers still perceive rubrics as having higher reliability and fairness. Though this research is particularly relevant in the context of architecture, it contributes to wider discussions about reliable and fair assessment of creative student assignments.

1 INTRODUCTION

1.1 Research aim

In higher education, the fairness and reliability of assessment with the purpose of student selection has always been a concern, due to the impact on students' chances in life. Especially in recent years, the use of high-stake assessments has increased in both scale and range (Stobart and Eggen 2012, 1). Universities continue to search for selection methods that are reliable, time efficient, and ensure equal chances for students of dispersed backgrounds, which is no easy task, especially when the judgement of human assessors is involved.

To select students for admission, prospective students complete an assignment to show their mastery of the required skills and competencies. It is common in education to quantify the quality of the assignment into grades, often through the use of rubrics. Such a way of grading student assignments by one or two teachers has several downsides. First, assessment tasks standardised to ensure fairness, resulting in a suboptimal validity (Pollitt 2004, 4-5). Second, grading (more authentic) open-ended tasks requires a lot of time, as it is impossible to anticipate all answers that students could give and capture those in a comprehensive rubric. And, even with a rubric, the judgement underlying grading comes intuitively (Brooks 2012, 68). Third, the assessment becomes highly dependent on the judgement of one or two assessors, decreasing the reliability of grades due to the biases that assessors carry (Malouff and Thorsteinsson 2016, 249). A possible solution to these shortcomings, based on the assumption that people are better at using their professional judgement for comparing two assignments than reliably assigning a score to a single, isolated assignment, is pairwise comparative judgement (CJ). The current study aims to explore whether comparative judgement could serve as a fair and reliable method for student selection by comparing it to rubric-based assessment.

1.2 Research outline

In this study, we explore the fairness and reliability of CJ in the selection process of prospective students to the undergrad programme Architecture, Urbanism and Building Sciences (AUBS) at Delft University of Technology. Especially in this field, assessment is highly (inter)subjective. This case study compares the official procedure of selecting prospective students for assessment with the use of a rubric to a pilot assessment with CJ. We first review the theoretical background for CJ. Then, section 3 explains our mixed methods approach. In section 4 and 5 we present and discuss the results of our exploration, as well as the potential of CJ as an assessment method to be used in the context of architecture and beyond.

2 THEORETICAL BACKGROUND

2.1 Construct validity of comparative judgement

Substantiation for the appropriateness of using CJ in education can be found in existing literature. For instance, in a secondary education setting, a high construct validity was achieved with CJ. One of the main findings was that CJ privileged scientific understanding, whereas conventional grading methods favoured recall of facts. This finding supports the notion that CJ is suitable for assessing higher order skills. In the same study, however, the downside of CJ was found to be an increased rather than reduced workload for assessors (McMahon and Jones 2015, 380-2).

A similar study showed there being multiple 'types' of assessors, but differences in rankings compiled by these distinct types of assessors are small. All of the four types emphasised argumentation and structure in their judgement of academic writing, confirming the construct validity of the resulting ranking (Lesterhuis et al. 2022, 127-9). In practice this means that increasing the number of assessors adds extra criteria taken into account in assessment, and minimises the impact of one assessor placing a disproportional weight on a single criterion. Such a mechanism still occurs when assessors have different definitions of 'academic writing' (Van Daal et al. 2019, 70).

2.2 Reliability of adaptive comparative judgement

Comparative judgement has been around since the 1920's, when it was coined as a way to measure psychological phenomena. Its application in education remained uncommon until computers became readily available to enable the use of (adaptive) algorithms, which have decreased the workload previously associated with CJ (Pollitt 2012a, 159). The resulting method of adaptive comparative judgement (ACJ) is described as a scoring instrument involving decisions on the relative quality of students' work through pairwise comparisons, which are configured into a ranking by an adaptive algorithm (Pollitt 2012b, 283-4). Algorithms used for ACJ select the most informative comparisons to reduce the total number of comparisons needed.

Despite the reduction in the amount of comparisons needed, recent studies have concluded that the scale separation reliability (SSR) of adaptive algorithms tends to be exaggerated (Crompvoets, Béguin and Sijstsma 2020, 336; Kimbell 2022, 1523-4). Moreover, in terms of validity, the ACJ does not outperform non-adaptive methods of CJ (Bramley and Vitello 2019, 52). In a response to these findings, Crompvoets, Béguin and Sijstsma (2020) have developed an adaptive algorithm that takes into account the uncertainty of parameters related to the works being compared. Although the algorithm reduced the standard error, it still required 20 pairwise comparisons per student work to achieve a SSR of .80, equal to the number of comparisons appropriate for non-adaptive CJ. It is therefore necessary to measure the reliability of ACJ in other ways. Suggestions are to use correlations with relevant external variables, or compare different sets of assessors (Bramley 2015, 15). The current study proposes a third alternative by comparing the ranking compiled through ACJ to a ranking by the same group of assessors using a rubric.

2.3 Perceived fairness of comparative judgement

Several studies have found that students perceive comparative judgement as being more fair, as multiple assessors evaluate their work. The perspective of the assessor is less well researched. In one study, assessors considered the ranking to be informative, and were generally curious to see whether a work they view as being of good quality indeed gets placed high up in the ranking (Van Gasse et al. 2017, 12-3).

3 METHODOLOGY

This case study makes use of a mixed methods design, where the comparison of the two assessment methods in terms of reliability and time efficiency was assessed using quantitative methods, and fairness was evaluated qualitatively.

3.1 Data collection

Data collection consisted of several steps. First, the official, rubric-based selection process was carried out with 820 prospective students. Each student submitted an assignment consisting of 2 drawings and a supporting text, which was assessed by a pair of teachers from a group of 43 in total. All pairs consisted of one teacher from the Architecture department and one from the Urbanism department. In cases where the difference between the two teachers was equal to or higher than 9 points out of 30, the coordinator of the undergrad programme stepped in as a third assessor. In total, 450 students were admitted (see the left column of *Fig. 1*).

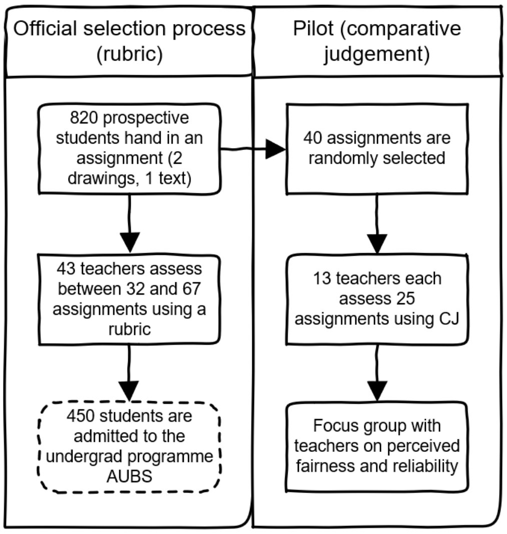
Official selection process (rubric)
Pilot (comparative judgement)
820 prospective students hand in an assignment (2 drawings, 1 text)
40 assignments are randomly selected
43 teachers assess between 32 and 67 assignments using a rubric
13 teachers each assess 25 assignments using CJ
450 students are admitted to the undergrad programme AUBS
Focus group with teachers on perceived fairness and reliability

Fig. 1. Schematic overview of the selection process and data collection for the pilot.

Second, as the right column of *Fig 1.* shows, the comparative judgement was conducted as a pilot experiment with a random sample of 40 assignments selected from the pool of 820 using a random number generator. 21 of these 40 assignments were handed in by students that were admitted to the undergrad programme, and 19 belonged to students not admitted. Out of the 43 assessors involved in the rubric-based selection, 13 signed up voluntarily to each compare 25 pairs of assignments. 2 of them were from the Architecture department; 11 were from Urbanism.

Finally, fairness was evaluated in a focus group session. The assessors provided a first impression of the CJ tool and its perceived reliability and fairness anonymously. The results were then discussed in plenary. Next, the two rankings were shown next to each other and the assessors were asked which they perceived as more reliable. They did not know which raking resulted from either rubric-based or CJ assessment. The input from the focus group was summarised based on notes and observations.

3.2 Data analysis

To determine the inter-rater reliability of the rubric-based assessment, the intraclass correlation coefficient (icc) was calculated. For CJ, the subset of 40 random assignments was uploaded by the researchers into the tool ‘Comproved’. The minimum number of comparisons per assessor needed to receive a reliability of .7, which is the suggested minimum for summative assessments, was calculated by using *Eq.1* (Goossens 2019, 12).

$$Comparisons\ per\ assessor = \frac{nAssignments\ \cdot 7{,}5}{nAssessors}$$

Eq. 1.

The algorithm in ‘Comproved’ determined which comparisons were shown to assessors. ‘Comproved’ also logged the time investment of each assessor, and displayed the average time per comparison. After each comparison, the tool recalculated the ranking of the assignments. When all comparisons were completed, a final ranking was produced together with its reliability coefficient.

To determine the overlap between the rankings, we computed Spearman’s rank-order correlation. The same was done to determine the difference in time investment across the two methods, where the time investment for the rubric-based assessment was based on an estimation provided by the involved assessors. The time investments were grouped in cohorts of 0-4, 4-8, 8-12, etc. hours to deal with the potential inaccuracy of the estimations. For the focus group, thematic analysis was used to draw conclusions about perceived fairness and reliability.

4 RESULTS

4.1 Quantitative outcomes

Each of the 820 assignments in the official selection process was assessed by two teachers (*icc* = .46, *p* < .001), between whom the inter-rater reliability appeared as poor. For the 69 out of 820 assignments were a third assessor stepped in, the reliability was especially low (icc = -.07, p = .09).

Each assignment in the CJ subset was assessed 14 to 16 times using an adaptive algorithm. Using Spearman’s rank-order correlation, the null-hypothesis is rejected, for $r_s(40) = .60, p < .001.$ This means there is significant overlap with the rubric-based ranking. However, when using a split-file Spearman’s rank-order correlation separating the students who are admitted under the rubric-based assessment from those not admitted, no relationship seems to exist between the rankings resulting from the two different assessment methods ($r_s(21) = .22, p = .36$; $r_s(19) = .37, p = .11$, for students with rank 1-21 and 22-40, respectively). The reliability of the CJ ranking is .61, based on the SSR calculated by ‘Comproved’.

Fig. 2 shows the two ranks per assignment. The CJ-based ranking differed from the rubric-based ranking to the extent that 4 out of 21 students admitted to the undergrad programme AUBS would not have been admitted if CJ was used for the selection process, and vice versa, 4 out of 19 would have been admitted with CJ.

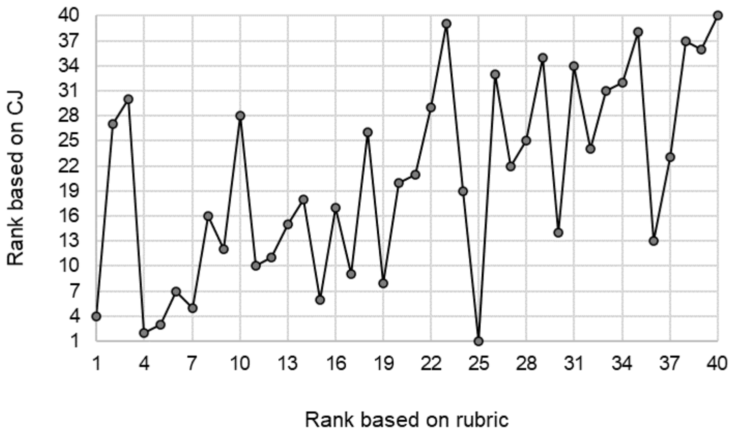

Fig. 2. Plot of each assignment’s rank according to the two assessment methods.

Concerning time efficiency, assessors reported spending between 1 and 12 hours on 32 up to 67 rubric-based assessments. The average time spent on assessing 25 assignments with CJ was 27 minutes (*min = 6,4 minutes*, *max = 121,1 minutes*).

4.2 Qualitative outcomes

The opinions of the teachers concerning reliability of the different assessment methods varied. The pilot with CJ made all participating teachers question the reliability of the rubric-based assessment. Still, most regarded rubrics as more reliable than CJ, because rubrics allow for the quantification of judgement instead of basing judgement on an overall feeling. In other words, rubrics were generally considered more objective and therefore also more reliable. On top of that, many assessors did not read the supporting text in the CJ assessment and based their choice on drawings only. Because the drawings concerned different buildings, it was also difficult to compare those. Teachers who preferred CJ in terms of reliability based their opinion on the higher number of assessors per assignment in CJ, and their observation that the rubric still left too much room for interpretation.

Regarding perceived fairness, there was again a preference for rubric-based assessment, as assessors took more time for each individual assignment using that method. CJ was found to be impersonal and did not stimulate the assessors to think deeper about whether the assignment reflected a capable and suitable student. Another reason why CJ was perceived as less fair was the absence of an option to rank two assignments as equal in quality, making the ranking somewhat arbitrary.

5 SUMMARY AND ACKNOWLEDGMENTS

5.1 Discussion

Before this study, there was little to no data available to provide insight into the reliability of high-stake assessment. The use of the tool ‘Comproved’ allows for drawing quantitative conclusions on the ranking’s reliability, whereas the reliability of rubrics is often not calculated. Having insight into an assessment’s reliability allows for taking measures to increase it (such as adding more assessors to the pool, or increasing the number of comparisons per assessors in the case of CJ).

This study shows a higher reliability for CJ compared to rubric-based assignment, albeit based on different measures of reliability. Additionally, the use of CJ could lead to a reduction in workload by allowing to assess more assignments in the same amount of time. In short, because the selection process already results in a ranking, CJ could make the process more efficient. However, most teachers still perceive assessment with rubrics as more reliable and fair. It therefore difficult to conclude based on this study that CJ improves the fairness of the selection process.

5.2 Limitations

First and foremost, the limited number of participating teachers prevented us from being able to replicate the entire ranking of 820 assignments with CJ. Involving more assessors would also allow for reaching a higher reliability. In addition, the skewed division over the Architecture and Urbanism departments meant that disproportional weight may have been attached to aspects concerning urbanism in the CJ ranking. The limited availability of the teachers also caused some of them to disregard the texts and only focus on the drawings that were part of the assignment, which partly explains difference between the two rankings.

Concerning the use of the tool 'Comproved', one assessor clicked on the wrong assignment once, which could not be undone. For high-stake assessments with big student groups such as the undergrad selection process, the option to mark two assignments as equal would have been beneficial for the (perceived) reliability, too.

5.3 Suggestions for future research

To decrease the workload of teachers, it would be interesting to incorporate more senior students into the pool of assessors, as a meta-analysis found no difference in the amount of comparisons needed to reach a reliable ranking when the assessors are teachers versus peers (Verhavert et al., 2019, 555).

It would also be worthwhile to ask students about the perceived fairness of both methods, in addition to asking teachers. Alternatively, if CJ were to be implemented in the selection process, the number of appeals lodged against admission decisions could be compared across years to give an indication of perceived fairness.

Lastly, with a larger sample of assignments, it would be interesting to investigate the tipping point of admission versus refusal: which assessment method is most reliable in distinguishing between students who are just above and below the tipping point?

5.4 Conclusion

To summarise, using an adaptive comparative judgement tool such as 'Comproved' could increase the reliability and efficiency of high-stakes assessment. Using a rubric, inter-rater reliability appeared to be low, and the assessment was highly time consuming. Yet, concluding if CJ is also a more fair alternative to using a rubric depends on the perception of all of us involved in architecture education. This study calls for further research on student views on high-stakes assessment to arrive at a method that satisfies universities' needs in ensuring equal chances for admission in the most reliable and time efficient way.

REFERENCES

[1] Stobart, G. and Eggen, T., (2012), High-stakes testing–value, fairness and consequences, *Assessment in Education: Principles, Policy & Practice*, Vol. 19, No. 1, pp. 1-6, DOI: 10.1080/0969594X.2012.639191.

[2] Pollitt, A., (2004), Let's stop marking exams, International Association of Educational Assessment Conference, Philadelphia, pp. 1-21.

[3] Brooks, V., (2012), Marking as judgment. *Research Papers in Education*, Vol. 27, No. 1, pp. 63-80, DOI: 10.1080/02671520903331008.

[4] Malouff, J. M. and Thorsteinsson, E. B., (2016), Bias in grading: A meta-analysis of experimental research findings, *Australian Journal of Education*, Vol. 60, No. 3, pp. 245-256, DOI: 10.1177/0004944116664618.

[5] McMahon, S. and Jones, I., (2015), A comparative judgement approach to teacher assessment, *Assessment in Education: Principles, Policy & Practice*, Vol. 22, No. 3, pp. 368-389, DOI: 10.1080/0969594X.2014.978839.

[6] Lesterhuis, M., Bouwer, R., Van Daal, T., Donche, V. and De Maeyer, S., (2022), "Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts", In *Frontiers in Education*, Vol. 7, pp. 122-131, DOI: 10.3389/feduc.2022.823895.

[7] Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V. and De Maeyer, S. (2019). Validity of comparative judgement to assess academic writing: Examining implications of its holistic character and building on a shared consensus, *Assessment in Education: Principles, Policy & Practice*, Vol. 26, No. 1, pp. 59-74, DOI: 10.1080/0969594X.2016.1253542.

[8] Pollitt, A., 2012a, Comparative judgement for assessment, *International Journal of Technology and Design Education*, Vol. 22, No. 2, pp. 157-170, DOI: 10.1007/s10798-011-9189-x.

[9] Pollitt, A, 2012b, The method of adaptive comparative judgement, *Assessment in Education: Principles, Policy & Practice*, Vol. 19, No. 3, pp. 281-300, DOI: 10.1080/0969594X.2012.665354.

[10] Crompvoets, E. A. V., Béguin, A. A. and Sijtsma, K., 2020, Adaptive Pairwise Comparison for Educational Measurement, *Journal of Educational and Behavioral Statistics*, Vol. 45, No. 3, pp. 316-338, DOI: 10.3102/1076998619890589.

[11] Kimbell, R., 2022, Examining the reliability of Adaptive Comparative Judgement (ACJ) as an assessment tool in educational settings, *International Journal of Technology and Design Education*, Vol. 32, pp. 1515-1529, DOI: 10.1007/s10798-021-09654-w.

[12] Bramley, T. and Vitello, S., 2019, The effect of adaptivity on the reliability coefficient in adaptive comparative judgement, *Assessment in Education: Principles, Policy & Practice*, Vol. 26, No. 1, pp. 43-58, DOI: 10.1080/0969594X.2017.1418734.

[13] Bramley, T. (2015). Investigating the reliability of Adaptive Comparative Judgment. Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

[14] Van Gasse, R., Mortier, A., Goossens, M., Vanhoof, J., Van Petegem, P., Vlerick, P. and De Maeyer, S., 2017, "Feedback opportunities of comparative judgement: An overview of possible features and acceptance at different user

levels", In *Technology Enhanced Assessment*: *19th International Conference, TEA 2016, Revised Selected Papers*, Vol. 19, pp. 23-38, Springer International Publishing.

[15] Goossens, M., 2019, "Comproved: Practical guide for instructors", https://comproved.com/wp-content/uploads/2019/12/Practical-guide-instructors_ENG_12_2019.pdf.

[16] Verhavert, S., Brouwer, R., Donche, V. and De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice,* Vol. 26, No. 5, pp. 541-562. DOI: https://doi.org/10.1080/0969594X.2019.1602027