ASSESSMENT OF COMPETENCY DEVELOPMENT IN A CHALLENGE-BASED LEARNING COURSE: CAN COACHES BE OBJECTIVE ASSESSORS?

N. Petrová
University of Twente
Enschede, The Netherlands
0000-0003-0256-6289

L. Chapel[1]
University of Twente
Enschede, The Netherlands
0000-0003-2319-8737

L. G. A. Buunk
University of Twente
Enschede, The Netherlands
0000-0002-6716-8778

G. H. Kaptijn
University of Twente
Enschede, The Netherlands

ABSTRACT

Higher education institutions aim to incorporate competency development into their engineering curricula, which can help engineering students become independent critical thinkers with entrepreneurial mindsets. However, no solid methods exist to evaluate the acquisition of these competencies. Such assessments' objectivities are often ensured by distinguishing between who supervises a student group and who grades its project. The assessor's active involvement in the learning process is essential for assessing competency development during the learning process, but such involvement may lead to assessor bias. This study aims to investigate whether and under what conditions coaches can be objective assessors. An intraclass correlation coefficient (ICC) was used to measure the level of agreement between assessors and coaches when using the same rubric to assess students' deliverables. Four assessors and seven coaches from the University of Twente assessed 24 students' individual learning processes based on individual reflection deliverables. The coaches assessed the students they supervised during a challenge-based learning (CBL) course, while

[1] *Corresponding Author*
Leonie Chapel
l.bosch-chapel@utwente.nl

the assessors were without participating in the learning process assigned randomly to students. The means were compared using SPSS, which indicated, among other things, that coaches generally awarded higher scores than assessors. This may indicate that coaches are biased because of their involvement in the learning process. Despite this, the results also indicate that coach assessment was in line with assessors when the coach was an appointed and experienced examiner.

1. INTRODUCTION

Higher education institutions aim to better prepare their students for the labour market by facilitating their development of transferable skills and lifelong learning competencies [1], [2]. The inclusion of competency development in higher education calls for new assessment methods, and one of the most promising and innovative approaches that can support such a transition to futureproof higher education is challenge-based learning (CBL) [3]. Challenge-based learning builds on experiential learning theories that view the learning process as being more important than the learning outcomes. Furthermore, these theories argue that skills are learned best in environments that resemble real-life situations [4].

This research was conducted on the Autumn Challenge programme, which is an extracurricular, international CBL programme organised by the University of Twente between October 2021 and January 2022. In this online programme, students from four ECIU[2] member universities worked on seven different challenges that resembled real-life situations and were under the theme of UN SDG 11. The programme was open to second- and third-year bachelor's and master's students from all ECIU universities.

1.1 Assessing competency development

Higher education institutions are increasingly exploring whether students can learn pass tests but also to gain a deeper understanding of the knowledge. The difference between knowledge assessments and assessments that focus on the learning process is that the latter also enables students to develop further after examination. Even though certain difficulties exist when assessing competency development, benefits exist as well. For instance, assessment drives learners to develop transferable skills [5], but also generates a higher level of commitment [6], thereby leading to more motivated students. Feedback also plays a key role, as does the active role that students play in their learning processes. To be able to assess students' learning processes, assessors should play an active role in these processes as well.

An important requirement for competency assessment is objectivity, which is often ensured by distinguishing between those who supervise students and those who assess their final outcomes [7]. In a CBL course, this means that those who coach students should not play a role in assessing the students' deliverables and vice versa [8]. However, due to the focus on the learning process in CBL, the assessor's active involvement during the learning process may provide valuable insights in addition to the deliverables on which the students can be assessed [7], [9]. Thus, combining the assessor and coach roles may provide additional insights into students' learning processes, but also can elicit assessor bias. Such a bias, whether positive or negative,

[2] https://www.utwente.nl/en/eciu/

can cause errors in assessor judgements, eliciting the potential to negatively impact an assessment's objectivity and quality [10].

1.2 On (perceived) biases in the case of the Autumn Challenge programme

An objective assessor is one who assesses the students' deliverables based solely on the criteria of the relevant assessment rubrics and the weighing thereof. However, when subjective factors (partially) influence an assessment, the assessment is 'biased'. Subjective factors can surface in many forms, e.g. the assessor's impression of a student, considering factors that are not part of the assessment rubric, or focusing too much on parts of the assessment rubric that are closest to the assessor's interest(s) or perspective(s) [8]. Various types of grading biases are recognised in the literature [8] and are summarised in Table 1. Due to the coaches' involvement in the learning process in the Autumn Challenge programme, there were five types of bias with increased potential: the contamination effect*,* halo effect, horn effect, norm shift and signal effect.

Table 1: Overview of the most common types of grading biases [8]

Type of bias	Information
Contamination effect	The contamination effect is the effect that occurs in grading when the freedom in grading, involuntary or random, is used for purposes other than those of an uninhibited, unbiased assessment. This is the case, for example, when assessors give lower scores to show that their subject is difficult.
Halo effect	The halo effect occurs when assessors allow their judgement to be influenced by other performances of the student than those expressed in the performance to be assessed. In this case, assessors tend to judge a 'good' student's performance as somewhat higher than warranted by the student's performance. The performance is overvalued.
Horn effect	The horn effect is the opposite of the halo effect. The Horn effect occurs when assessors allow their judgement to be influenced by the student's performance other than expressed in the performance to be assessed. In that case the assessors tend to assess a 'bad' student at a somewhat higher performance level than the performance of the student justifies. The performance is undervalued.
Norm shift	The norm shift is the effect that occurs when an assessor adjusts to the performance of students. For example, an appraiser may become less harsh if after a number of assessments, it is found that most students answer the same question incorrectly.
Restriction of range	Restriction-of-range is caused by the freedom in the assessment task that willy-nilly leads to certain distributions of the ratings that express general human or personal tendencies. For example, one appraiser may use all scale values (scores from 1 to 10), while another will always assign values near the middle.
Sequence effect	The sequence effect occurs when an assessor incorrectly allows the assessment to be guided by one or more previous assessments. For example, assessors review test questions in a certain order. A large number of bad answers followed by a correct answer may then lead to the correct answer being graded disproportionately.
Signal effect	The signal effect occurs when assessors pay attention to different aspects, or when they weigh the aspects differently in their judgement. This effect occurs, for example, when assessors evaluate writing products. One assessor pays attention to grammatical (in)correctness and the other to structure or content.

1.3 Problem statement

In this paper, we use the term 'grading bias' to refer to unintentional grading bias. There is a considerable body of literature on how to prevent this, but these solutions are mainly aimed at eliminating situations where bias can be stirred up, such as intensive collaboration between lecturers and students. Is a certain degree of bias a real issue?

Research has shown that, when students have the feeling that they are not being judged for their work but, for example, for their personality or their past performance, both the student and their peers may feel that they have no influence on the outcome, no matter their efforts.

2. METHODOLOGY

This section of the paper outlines the methodology used to answer the following research question: Can coaches be objective assessors, and, if so, under which conditions?

2.1 Subjects

Student population

In total, 24 second- and third-year bachelor's and master's students participated in the programme in teams of three to five members each based on multidisciplinary, intercultural diversity and their preference for an overarching project. The teams' progress were monitored at weekly team coaching sessions as well as during three programme milestone moments.

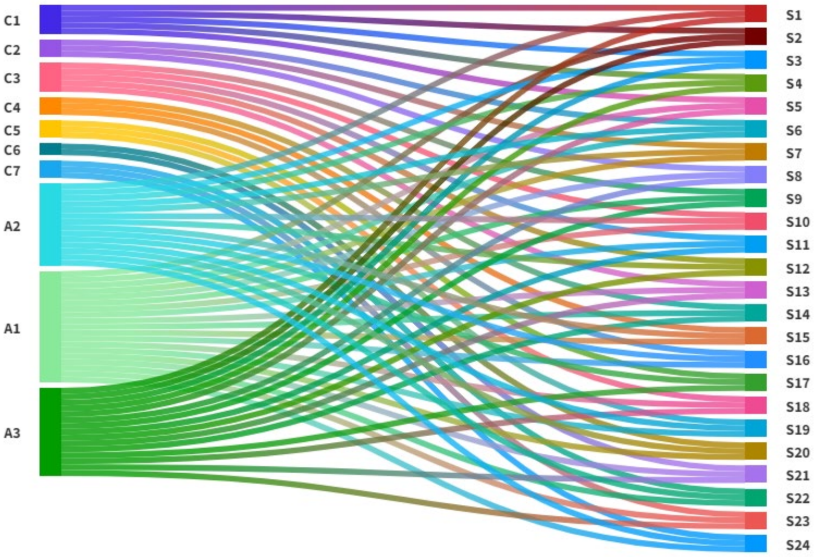
Challenge-based learning coaches

The seven student teams were each supported by their own team coaches. The coach's role in the learning process was to provide proper, flexible, and personal support to the students in their teams in order to enhance the learning process and team dynamics [11]. Two of these coaches were experienced academics and trainers with a University Teaching Qualification (UTQ) and experienced with CBL. All seven coaches received training on coaching in a CBL prior to the start of the programme. The coaches have been actively involved in the students' learning processes throughout the course, unlike the assessors.

Assessors

Just like the coaches, there were also four assessors that evaluated the students' final learning outcomes. Unlike the coaches, the assessors have not been involved in the students' learning processes. Each student has been assessed by his own coach and by two assessors. Details on who assessed the students can be found in Figure 1.

Figure 1: Distribution of the individual deliverables among coaches and assessors. A 'C' refers to a 'coach', an 'A' to an 'assessor' and a 'S' to a 'student'.



2.2 Assessment rubric

The assessment rubric for individual reflection was developed to provide students with the freedom to reflect in a way that suits them. Table 2 shows a summary of the assessment rubric (see Appendix A for the full rubric), which consisted of three levels of 'pass' scores (Excellent, Satisfactory, Sufficient) and one 'Insufficient' score for each assessment topic. For this analysis, each assessment scale received a corresponding score as shown in Table 2. Three assessment criteria were used to score and analyse five assessment topics, namely Professional growth, Skill development, Team role, Problem solving and Development of new skills and competencies. The 'report/video length' assessment criterium was excluded from the analysis, as no differences between the raters' assessments would be found.

Table 2: Overview of the assessment criteria used in the analysis

Assessment criteria		**Insufficient**	**Sufficient**	**Satisfactory**	**Excellent**
		2,5	5,5	7,5	9,5
Individual development	Professional growth	Reflects upon their educational and professional growths. Provides examples and reflects upon the BuddyCheck matrix related to individual development and programme components (Skills Labs, Thematic Lectures, Virtual Teams Group Work).			
	Skill development	Reflect upon the development of skills/competencies related to their individual learning goal set prior to the programme.			
Individual role	Team role	Reflects upon their role in the team. Provides examples.			
	Problem solving	Reflects upon their contribution to the problem-solving process. Provides examples.			
Individual effort	Development of new (interdisciplinary) skills and competencies	Reflects upon their effort to develop new skills and obtain new competencies in the programme. Provides examples and reflects upon the BuddyCheck matrix related to individual effort and programme components (Skills Labs, Thematic Lectures, Virtual Teams Group Work).			

2.3 Reliability of agreement

To ensure that the rubric provides sufficient consistency, and therefore any found differences are not likely to be caused by the interpretation of the assessment items but rather by the assessment itself, an intraclass correlation coefficient (ICC) was used. The outcomes were interpreted as follows: < 0.50, poor; between 0.50 and 0.75, fair; between 0.75 and 0.90, good; above 0.90, excellent. The ICC for intra-rater reliability for rater 1 and rater 2 was fair as it was 0.62 (0.25-0.83); for rater 1 and rater 3, was fair as it was 0.58 (0.19-0.81); and for rater 2 and rater 3, was excellent as it was 0.93 (0.82-.0.97). We can thus conclude that the rubric used ensured a high enough level of agreement between the raters to be able to compare the results.

2.4 Measures

In the case of the Autumn Challenge programme, the assessment structure is two-fold, i.e. assessors and student groups' coaches assess reflection deliverables. The assessors are not involved with the students in the course and, thus, can only base their assessments on the students' deliverables. To investigate this phenomenon further, perceived bias was measured.

Two outliers were detected that were more than 1.5 box lengths from the edge of the box in a boxplot. Inspection of their values did not reveal them to be extreme and they were kept in the analysis.

3. RESULTS

3.1 Data preparation

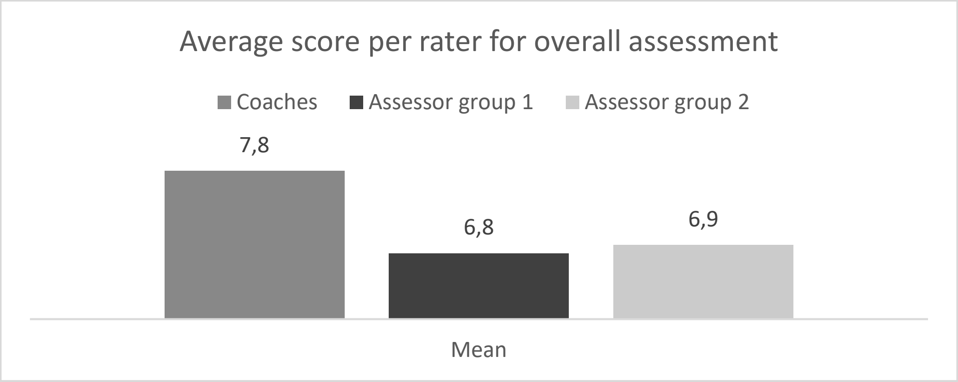
Before comparing the rater groups with one another, the first step was checking for the assumptions of normality and outliers. One outlier was found in the coach group. However, the coach in question is the most experienced teacher in that group, possesses the UTQ, and is very experienced in CBL education. Additionally, the outlier was not more than 1.5 box lengths from the edge of the boxplot. For these reasons, the outlier was not removed. In the first assessor group, outliers were detected, but both of them were no more than 1.5 box lengths from the edge of the box. Inspection of their values did not reveal them to be extreme and they were also kept in the analysis, as both outliers were detected in the overall score and the assessment criteria 'individual development'. In the second assessor group, there were no outliers. Furthermore, six individual cases were excluded from the analysis due to incomplete information, thus 18 cases were used.

3.2 Analysis

Assessment means comparison

The means of the assessment of the students' individual deliverables of the three raters (Coaches, Assessor group 1, Assessor group 2) are firstly compared as a whole as shown in Figure 2, and then on the three assessment criteria as shown in Table 3.

Figure 2: Means comparison of the three groups of raters for individual deliverable



Coaches score, on average (M = 7.8, SD = 1.096), much higher than the assessors, respectively (M = 6.8, SD = 1,535) (M = 6.9, SD = 1.944), who have not worked intensively with the students. Additionally, the average score for the 'individual development' assessment criterium given by the coaches was the highest (M = 7.8) followed by the second assessor group (M = 7.2) and the first assessor group (M = 7.0). Similarly, the highest average score for the 'individual role' assessment criterium was given by the coaches (M = 8.2), followed by the second assessor group (M = 7.2) and the first assessor group (R2 = 7.0). For the assessment criterium, 'individual effort', the highest average score was given by coaches (M = 7.3), which was followed by the second assessor group (M = 6.3) and the first assessor group (M = 6.2).

Looking at the individual assessment criteria, the differences are smallest for development and largest for role and effort. A higher assessment of 'individual role' and 'individual effort' by the coaches may indicate the influence of the halo or signal effects because, in CBL, coaches are part of the team and this may lead to an overvaluation of the students' roles and efforts in the project (*halo effect*). The coaches may also consider the role the students play and the effort they put into the project group to be more important than the students' individual development (*signal effect*).

Table 3: Means comparison of the three groups of raters for different assessment criteria. The highest scores are highlighted in bold.

		Individual Development		**Individual Role**		**Individual Effort**	
	N	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
Coaches	18	**7.8**	1.331	**8.2**	0.943	**7.3**	1.886
Assessor group 1	18	7.0	1.996	7.0	1.098	6.2	2.346
Assessor group 2	18	7.2	2.211	7.2	1.816	6.3	2.624

T-Test

Overall, the coaches scored statistically significantly higher on the overall assessment score compared to both assessor group 1 ($t(17) = 3.817$, $p < 0.001$) and assessor group 2 ($t(17) = 2.447$, $p < 0.026$). The coaches scored statistically significantly higher on the assessment criterium ‘individual role’ compared to assessor group 1 ($t(17) = 6.059$, $p < 0.001$) and assessor group 2 ($t(17) = 2.749$, $p < 0.014$). The coaches ($t(17) = 2.247$, $p > 0.38$) did not score statistically significantly higher on the assessment criterium ‘individual effort’ compared to assessor group 1 ($t(17) = 2.294$, $p > 0.35$). As expected, the assessment criterium ‘individual development’ was not significant compared to assessor group 1 ($t(17) = 1.750$, $p > 0.98$) and assessor group 2 ($t(17) = 1.225$, $p > 0.24$).

Pass/Fail comparison

As stated in the literature, an assessment that is not only based on visible and expected assessment but where other factors are also taken into account is undesirable [8]. This is the case regardless of whether this leads to a more positive or negative result. Apart from the fact that a grade should reflect the actual performance, there is an even greater risk that students who do not sufficiently master the learning objectives will still receive a ‘pass’ grade or vice versa. Table 4 shows that, based on the scores of the coaches, 10% of the students would have received a pass score without having objectively achieved the learning objectives.

Table 4: Differences between pass/fail scores given by the three groups of raters.

		Individual Development		Individual Role		Individual Effort	
	N	Pass	Fail	Pass	Fail	Pass	Fail
Coaches	18	17	1	18	0	17	1
Assessor group 1	18	14	4	18	0	15	3
Assessor group 2	18	14	4	16	2	14	4

It can, therefore, be concluded that coaches are less likely to fail students than assessors. The coaches may overvalue their students' performance (*halo effect*) because they may, to a certain extent, feel responsible for their individual development and the effort they put in.

4. CONCLUSION

Ensuring objectivity in competency assessment is one of the key requirements for valid assessment and future facilitation of transferable and lifelong learning competencies in higher education. The research data indicates that, in the case of the CBL Autumn Challenge programme, coaches were not always objective assessors. Coaches rewarded students whom they have coached with overall higher scores as compared with the assessors who were not involved in these students’ learning. Moreover, the coaches gave a fail to fewer students than the assessors did. Based on these outcomes, it can be concluded that bias influenced the coaches’ assessments. The coaches either tended to overvalue their students’ performances (*halo effect*) or give more weight to the role they played in the group work than to the other assessment criteria (*signal effect*). Nevertheless, our results also indicate that, when a coach was an experienced and trained teacher, disagreement between the assessment of the coach and the assessors was eliminated. This finding is also in line with Sa et al.'s [7] research, which concluded that more rigorous training for coaches in an open-ended project is required in order to assure valid assessment. However, because only two coaches in the Autumn Challenge programme were experienced and trained teachers, further research into the objectivity of experienced and trained coaches as objective assessors is recommended in order to further validate our outcomes.

REFERENCES

[1] R. . Klaassen, C. Milano, M. B. van Dijk, and R. . Bossen, "How to embed 'the reflective engineer' in higher engineering education," in *Proceedings SEFI 49th Annual Conference 2021*, 2021, pp. 968–973.

[2] J. Membrillo-Hernandez and R. Garcia-Garcia, "Challenge-Based Learning (CBL) in engineering: Which evaluation instruments are best suited to evaluate CBL experiences?," 2020. doi: 10.1109/EDUCON45650.2020.9125364.

[3] L. Chapel, N. Petrová, E. Tsigki, L. Buunk, and F. van den Berg, "Creating the conditions for an online challenge-based learning environment to enhance students' learning," in *Proceedings SEFI 49th Annual Conference 2021*, 2021, pp. 721–735.

[4] G. Mulgan, "Challenge-driven universities to solve global problems," *Nesta*, 2016.

[5] C. Hughes and S. Barrie, "Influences on the assessment of graduate attributes in higher education," *Assess. Eval. High. Educ.*, 2010, doi: 10.1080/02602930903221485.

[6] B. Clayton, K. Blom, D. Meyers, and A. Bateman, *Assessing and certifying generic skills What is happening in vocational education and training?* Adelaide, South Australia: National Centre for Vocational Education Research NCVER, 2003.

[7] T. Papinczak, L. Young, M. Groves, and M. Haynes, "An analysis of peer, self, and tutor assessment in problem-based learning tutorials," *Med. Teach.*, 2007, doi: 10.1080/01421590701294323.

[8] H. van Berkel, A. Bax, and D. Joosten-ten Brinke, Eds., *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleu van Loghum, 2014. doi: 10.1007/978-90-368-0239-0.

[9] B. Sa, C. Ezenwaka, K. Singh, S. Vuma, and M. A. A. Majumder, "Tutor assessment of PBL process: Does tutor variability affect objectivity and reliability?," *BMC Med. Educ.*, 2019, doi: 10.1186/s12909-019-1508-z.

[10] P. L. Hardré, "Checked Your Bias Lately? Reasons and Strategies for Rural Teachers to Self-Assess for Grading Bias," *Rural Educ.*, 2018, doi:

10.35608/ruraled.v35i2.352.

[11] M. MacLeod and J. T. van der Veen, “Scaffolding interdisciplinary project-based learning: a case study,” *Eur. J. Eng. Educ.*, 2020, doi: 10.1080/03043797.2019.1646210.

APPENDIX A

Assessment rubric: Individual Reflection

Criteria	Pass			Insufficient
	Excellent	Satisfactory	Sufficient	
INDIVIDUAL DEVELOPMENT	Reflects **extensively** upon their educational and professional growth. Provides **specific** examples and **fully** reflects upon the BuddyCheck matrix related to individual development and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **no** additional clarification is needed.	Reflects **adequately** upon their educational and proffesional growth. Provides **some** examples and reflects upon the BuddyCheck matrix related to individual development and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **some** additional clarification is needed.	Reflects **sufficiently** upon their educational and professional growth. Provides **few** examples and briefly reflects upon the BuddyCheck matrix related to individual development and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **substantial** additional clarification is needed.	**Absent** or **very limited** reflection upon their educational and professional growth
	Reflect **extensively** upon development of skills/competencies related to their individual learning goal that they set prior to the programme. Provides **specific** examples, so **no** additional clarification is needed.	Reflects **adequately** upon development of skills related to their individual learning goal that they set prior to the programme. Provides **some** examples, so **some** additional clarification is needed.	Reflect **sufficiently** upon development of skills related to their individual learning goal that they set prior to the programme. Provides **few** examples, so **substantial** additional clarification is needed.	**Absent** or **very limited** reflection upon upon development of skills related to their individual learning goal.
INDIVIDUAL ROLE	Reflects **extensively** upon their role in the team. Provides **specific** examples that support their reflections.	Reflects **adequately** upon their role in the team. Provides **some** examples that support their reflections.	Reflects **sufficiently** upon their role in the team. Provides **few** examples that support their reflections.	**Absent** or **very limited** reflection on their role.
	Reflects **extensively** upon their contribution to the problem solving process. Provides **specific** examples that support their contribution.	Reflects **adequately** upon their contribution to the problem solving process. Provides **some** examples that support their contribution.	Reflects **sufficiently** upon their contribution to the problem solving process. Provides **few** examples that support their contribution.	**Absent** or **very limited** reflection upon their contribution to the problem solving process.
INDIVIDUAL EFFORT	Reflects **extensively** upon their effort to develop new skills and obtain new competencies in the programme. Provides **specific** examples and **fully** reflects upon the BuddyCheck matrix related to individual effort and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **no** additional clarification is needed.	Reflects **adequately** upon their effort to develop new skills and obtain new competencies in the programme. Provides **some** examples and reflects upon the BuddyCheck matrix related to individual effort and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **some** additional clarification is needed.	Reflects **sufficiently** upon their effort to develop new skills and obtain new competencies in the programme. Provides **few** examples and **briefly** reflects upon the BuddyCheck matrix related to individual effort and programme components (Skills Labs, Thematic Lectures, Virtual Teams group work), so **substantial** additional clarification is needed.	**Absent** or **very limited** reflection on their effort.
REPORT/VIDEO LENGHT	Written report: The report lenght is **within** the given word count (600-650 words, excluding references) Video/animation: The video/animation is **within** the given time count (5-10 minutes)	Written report: The report lenght **slightly violates** the given word count (by <75 words, excluding references) Video/animation: The video/animation lenght **slightly violates** the given time (by <2 minutes)	Written report: The report lenght **violates** the given word counts (by >75 but <150 words, excluding references) Video/animation: The video/animation lenght **violates** the given time (by >2 but <5 minutes)	Written report: The report lenght is **largely outside** the given word count (by >150 words, excluding references) Video/animation: The video/animation lenght is **largely outside** the given time (by <5 minutes)