Automated Consensus Moderation as a Tool for Ensuring Reliability in a Multi-Marker Environment

Bertram Haskins, Nelson Mandela University, Port Elizabeth, South Africa, 6001

Students in programming subjects frequently have to submit assignments for assessment. Depending on the class size, these assignments may be divided amongst multiple, trained markers to be marked using a pre-defined rubric. Experience and differing opinions might yield different marks by different assessors on the same assessment. This yields an inconsistent marking process.

Consensus moderation is a technique whereby consensus is reached on a student assignment by incorporating the opinions of multiple markers. In this study, an automated form of consensus moderation is proposed in which the opinion of an individual marker on a specific criteria point on a rubric is cast as a vote. In this process a majority vote determines the successful completion of the specific rubric criteria point.

Tests are conducted in order to determine whether such an automated consensus moderation process yields more reliable results than that of individual markers. Using Krippendorf's α the average level of agreement between individual markers on 4 programming assignments is calculated as 0.522. This score is deemed unreliable. The individual markers show an average level of agreement of 0.811 with the automated consensus moderated result. This is classified as an acceptable level of reliability.

***Index Terms*—Education, Computer aided instruction, Programming**

I. Introduction

One of the main tasks required of academics is the marking of student assignments. In many cases marking is allocated to assistants in an attempt to alleviate the burden. In some cases, with large student group numbers, the marking is distributed between multiple markers, with each assignment only being marked once. This may lead to very different results in the interpretation of the marking rubric among the various markers. This has the effect of leading to inconsistent marking results, which may be queried when the results are returned to the students.

One possible solution to this problem is to apply consensus moderation. Consensus moderation is defined as tasking multiple markers to the same assignment in order to have them mark it independently. The individual markers then convene as a group in order to deliberate a final mark for an assignment based on consensus [1]. Using this approach would result in a lower workload than having an individual marker mark all the assignments, since the marker would only need to mark enough so that a majority vote can be applied. This study proposes an automated form of consensus by using marker rubric scores as votes. The main objective of this study is to determine whether this form of automated consensus moderation provides any benefit when applied to student assignments.

The remainder of this paper is structured as follows; Section II provides an overview of related work. Section III outlines the programming module on which this study is based, as well as the process employed in marking the assignments. Section IV contrasts the differences in marking results obtained from the 3 individual markers. Section V provides insights gained from the results and the study is concluded in Section VI.

II. Related Work

A. Consensus Moderation

In subjects where students need to complete assessments, there is a need for these assessments to be marked and in, in some cases, moderated. There are also cases in which there are multiple examiners and / or moderators. In general, moderation methods can be categorised as following either an inspection, statistical or consortium / consensus-based model [2]. An inspection-based model requires moderators to inspect a subset of teacher-marked work and adjusting the marks if necessary. A statistics-based model requires that assignment scores be adjusted according to a calculated measure.

Consensus moderation, which has also been referred to as social moderation, is a process which is frequently used by teachers to review and coordinate the ratings of student work [3]. This technique is frequently used for staff development as it helps new staff members realise how marking should be done. When applied to a set of papers from an individual teacher's classroom, the papers may be reviewed by teachers from the same school or by expert raters from other schools. Differences in ratings would be discussed in an effort to reach consensus. In order to judge the value of such a process, it is necessary to compare the qualifications of the judges and the (statistical) level of inter-judge agreement [3].

Consensus moderation is generally seen as a participative process in which assessors are required to meet at a pre-set location; apply their combined expertise and respect for each other's opinions to arrive at a final mark for an assignment under discussion. Although this process yields a more reliable result, since it is based on the opinion of more than one assessor, it does add to the time required for finalising the marking process and also requires that all the assessors find a common time and venue for a physical meeting. This process could be expanded to include ICT-based solutions such as meeting using an on-line meeting tool [4].

This study aims to expand upon this principle by substituting the consensus discussion by means of an on-line tool. In this process the opinion of every marker / expert will be represented by their marked version of an assignment. The assignments are marked by means of a rubric, with each item on the rubric being seen as an expert opinion; counting as a vote towards a final mark for the assignment.

B. Marking by Rubric

Deriving an opinion based upon various dictionary definitions, [5, p. 2] defines a (scoring) rubric as guidelines for scoring responses or criteria for assessing complicated things. Beyond providing guidelines for markers, a rubric provides details to students as to what is required of an assessment. Rubrics may be created in 3 main formats, namely, checklist, holistic or analytic, or in a fourth format which is a hybrid combination of the 3 [6]. The checklist format consists of a set of items which are either ticked as present / completed or not. There is no score allocated to a specific item. A holistic or performance level rubric is similar to a grading system, e.g. A to F, and provides a single description per item from which a marker then derives a grade. An analytic rubric provides a list of evaluation criteria / descriptions, with a number of points attributed to each item or criterion. These rubrics may be combined in various ways, such as creating a scored checklist by combining the concept of the checklist with an analytic rubric.

When setting a rubric it is necessary to first decide whether the rubric will be used solely by the markers or also by the students for assessing the completeness of their work before submission. By providing the rubric before submission it may ensure that students perform every task required by the rubric, but might also inhibit any creativity on the part of the student. Figure 1 presents an overview of the process required for setting an appropriate rubric. The figure is an abbreviated adaptation of the general steps provided by [6].

The first step in the process is to determine the task which requires rubric assessment. This could be the specifications or instructions for an assignment. The next step is to determine which kind of rubric is required, i.e. checklist, holistic, analytic or a hybrid approach. It is recommended that before setting a new rubric an investigation is done to see if there are any existing general-purpose or previously used rubrics which could serve as a framework for the new rubric. [7] discusses multiple scenarios in which rubrics devised for one task may be adapted to serve another. The *Devise Criteria* step involves setting the individual items associated with the rubric. The items should be sequenced logically and could be grouped, if necessary. During this step, the layout format of the rubric items should be defined; as well as their descriptive labels. [8] suggests that when criteria are defined, the examiner should consider whether the scoring categories are well-defined, the differences between the categories are clear and whether two independent raters would arrive at the same score when using the rubric. This might not always be possible if the work being assessed is very subjective, but the examiner should strive for rubric clarity as far as possible.

Once set, the rubric should be tested against either an actual student submission or a hypothetical set of best and worst case assessments. The results from testing should be evaluated to determine if the rubric is suitable for performing a complete assessment. If the rubric does not clearly delineate performance levels, the criteria may need to be revised and the process repeated. Once the rubric is deemed suitable it may be implemented to perform an assessment. This entire process speaks to the validity of the rubric, but there is also the aspect of reliability to consider [8].

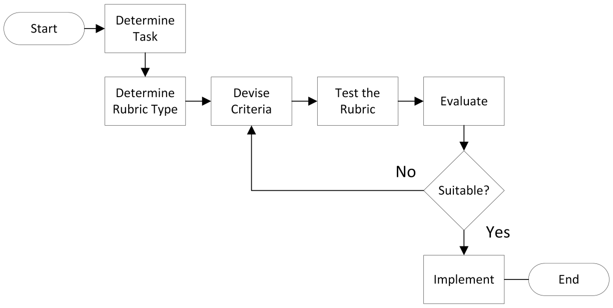


Fig. 1. The Rubric Creation Process

C. Evaluating Reliability

Although care may be taken in the creation of the rubric, it still needs to be applied by a human evaluator. Human evaluators can have differing opinions on the allocation of marks. When the evaluators have a part in setting the rubric, there might be less of a difference in opinion. However, when a single examiner sets the rubric, and requires other evaluators to apply the rubric, there might be a difference in opinion. Two indicators for assessing the reliability of the assessment process, when using a rubric, are inter-rater reliability and intra-rater reliability [8].

Intra-rater reliability refers to external factors influencing the marking consistency of a single marker. This may lead to differences in mark allocation when the marker is in a different emotional state. This might be difficult to quantify. A simpler measure of whether a rubric is reliable is by means of inter-rater reliability or agreement.

Inter-rater reliability can be seen as a measure of the level of agreement between two markers or raters. There are a few widely used means of calculating inter-rater agreement. Two of these are Cohen's κ and Krippendorff's α.

Cohen's κ is a statistical calculation for summarising the cross-classification of two nominal variables [9]. The calculated value is in the range of -1 to 1. Total agreement between coders (markers or raters) is signified as 1. Zero signifies the expected result under total independence and a negative value occurs when the calculated agreement is less than what would be expected under chance conditions. Cohen's κ is calculated as follows:

$$\kappa = \frac{P - E}{1 - E} \tag{1}$$

In this equation, *P* refers to the relative observed agreement between markers and *E* refers to the hypothetical probability of chance agreement. [10] and [11] both make use of Cohen's κ to measure agreement between human participants.

Krippendorff's α is a very general measure of inter-rater agreement, which allows for uniform reliability standards to be applied to a great diversity of data [12, p. 221]. It is a measure of the extent to which the proportion of the differences in opinion between observers that are in error, deviates from the perfect agreement. This method is applicable to any number of values per variable, to any number of observers and sample sizes, various metrics and even data with missing values. In its most general form, it may be defined as shown in Equation 2

$$\alpha = 1 - \frac{D_o}{D_e} \tag{2}$$

In this equation D_o refers to the measure of the observed marker disagreement and D_e is a measure of the disagreement that can be expected when chance prevails. A calculated value of 1 represents perfect agreement, 0 may occur when observed and expected disagreements are equal. The reliability could, however, be measured as low as -1. A negative value may, however, be too far removed from the desired result of 1, to be of use. Lower (and negative) values may be the result of sampling errors and systematic disagreements. Sampling errors may occur when samples are too small, which may result in too few required observations. Systematic disagreements tend to happen when a marker misinterprets the rubric instructions.

III. Marking Process

The subject on which the practical observations of this study was done is a fourth year programming subject in an information technology course. The programming language used in the subject is Microsoft Visual C#. The course content focuses on software design patterns. The practical observations were conducted over 4 assignments. Each assignment had to be completed by students as individual work, i.e. no group assignment submissions were allowed. The assignments were each submitted as a single file, using a template provided to the students.

Three markers were appointed to mark the submitted assignments. All three of the markers were postgraduate students working towards a master's degree in information technology. One of the three markers has industry experience as a software developer. Two of the three markers were previously enrolled for a degree course in mechatronic engineering, which included software development subjects in their first and third years. These two markers are at least 2 years older than the third marker, with more experience in software development.

All 3 markers had previously been enrolled for the course on which the study is based; as well as having acted as assistants in practical lab sessions for these subjects. They are very familiar with the subject content. To ensure that the markers were well informed, they were individually briefed as to how to perform the marking. This briefing included a demonstration as to how one of the assignments was marked by the main examiner. They were also each provided with a set of guidelines as to how to apply the rubrics. Before embarking on marking, they were each requested to mark a few assignments as practice and then discuss their experience with the main examiner in order to improve their understanding of the marking process. These practice assignments were taken from a previous year's assignments and did not match the exact content and rubric items of any of the current assignments.

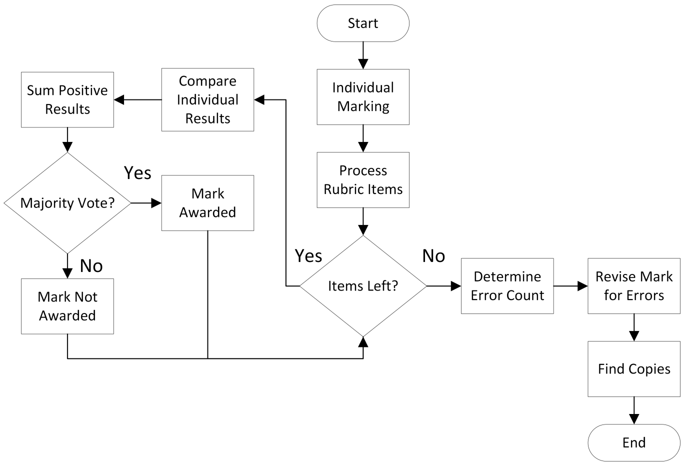


Fig. 2. The Automated Consensus Moderation Process

The three markers were required to mark the submitted assignments via a web-based interface. The marking interface was divided into 3 main sections, namely an itemised rubric, a section for capturing errors and a section for identifying copied assignments. Examples of these sections may be seen in Figures 3 to 5. All assignments start with a mark of a 100% and marks are deducted for non-adherance to the rubric or for errors encountered. The rubric focuses on the assignment specification requirements and as such does not take into account errors, such as code syntax errors. Syntax errors are captured using the interface shown in Figure 4. A 10% mark deduction is made for each error indicated by a marker. The markers were also provided with an area for indicating whether they thought a particular assignment was copied from another student (Figure 5). For the purposes of creating a complete dataset, every assignment was marked by each of the 3 markers.

The submitted assignments consist solely of code, but because of the nature of the assignments, students may have different approaches to incorporating the design patterns into their solutions. The assignment specifications were however devised to require very specific items, named in a specific fashion, in specific places in the code. This allowed the rubrics to be set so that the markers could simply check whether these items are in their specified locations and named correctly. A single mark would be allocated for each correctly placed item.

The values captured from the assignment rubrics were stored in a central Microsoft SQL Server database. Having the results in a central database allowed queries to be executed in order to retrieve the results presented in the next section.

By querying the database, results were retrieved for individual markers per assignments, comparisons between markers on assignments, consensus results for all markers per assignment and also comparisons of the results of individual markers with those of the consensus results. At a high level, the consensus process is laid out in Figure 2.

The process requires that for each assignment, the completed rubrics for the individual markers be processed. Whenever a marker marked a specific rubric criteria point / item as complete or present it is taken as a vote by that marker for that criteria point. If a majority of markers vote positive for the criteria point, it is awarded. The markers could also indicate a number of errors by means of a process separate

TABLE I
Inter-rater Reliability Between Rater Pairs Calculated as Krippendorf's α

Assignment	Marker 1 and 2	Marker 1 and 3	Marker 2 and 3	Average
Assignment 1	0.626	0.344	0.546	0.505
Assignment 2	0.847	0.547	0.501	0.632
Assignment 3	0.894	0.424	-0.044	0.425
Assignment 4	0.737	0.463	0.371	0.524
Averages	0.776	0.445	0.344	0.522

TABLE II
Comparison of the Percentage of Differences per Assignment Between Pairs of Markers

Assignment	Marker 1 and 2	Marker 1 and 3	Marker 2 and 3
Assignment 1	16	10	22
Assignment 2	17	12	6
Assignment 3	5	3	0
Assignment 4	13	30	17

TABLE III
Comparison of the Percentage of Differences per Assignment Rubric Item Between Pairs of Markers

Assignment	Items	Marker 1 and 2	Marker 1 and 3	Marker 2 and 3
Assignment 1	3404	11	17	14
Assignment 2	4368	35	39	46
Assignment 3	4650	22	39	43
Assignment 4	2300	11	12	13

to the rubric. In the case of the errors, the lowest common denominator number of errors indicated by each marker is taken as the final number of errors on the assignment. A mark deduction is then performed on the marked rubric according to the number of calculated errors. A similar process is applied to determine if the assignment might have been copied.

IV. Results

The first step in determining whether there is any need for consensus moderation (automated or not) is to determine whether there is any difference between the marking results presented by the various markers. According to [8], inter-rater reliability is a means of determining how close the relationship is between the marking performed by different assessors. For this study, Krippendorf's α was used calculate the level of inter-rater reliability (agreement). These calculations were done between pairs of markers, e.g. marker 1 and 2 or marker 1 and 3. The results of these calculations are shown in Table I.

From the results in Table I it would seem as if the level of agreement is very low; even after training the markers on the use of the rubric. The reason for this might be the multitude of approaches available to students for completing an assignment. On average, they have a level of agreement calculated as 0.776. With regards to agreement, [12, p. 241] suggests that only α values higher than 0.8 should be accepted as a measure of reliability. In all instances it would seem that marker 1 and 2 are in relatively high agreement, but both marker 1 and 2 have low levels of agreement with marker 3. This indicates that the third marker has a high level of disagreement with the other markers. This might be because the third marker misinterpreted or alternatively interpreted the rubric instructions. The alternative interpretation may have been the result of a different undergraduate programming experience to the other 2 markers, which have a shared mechatronic (as well as information technology) undergraduate background.

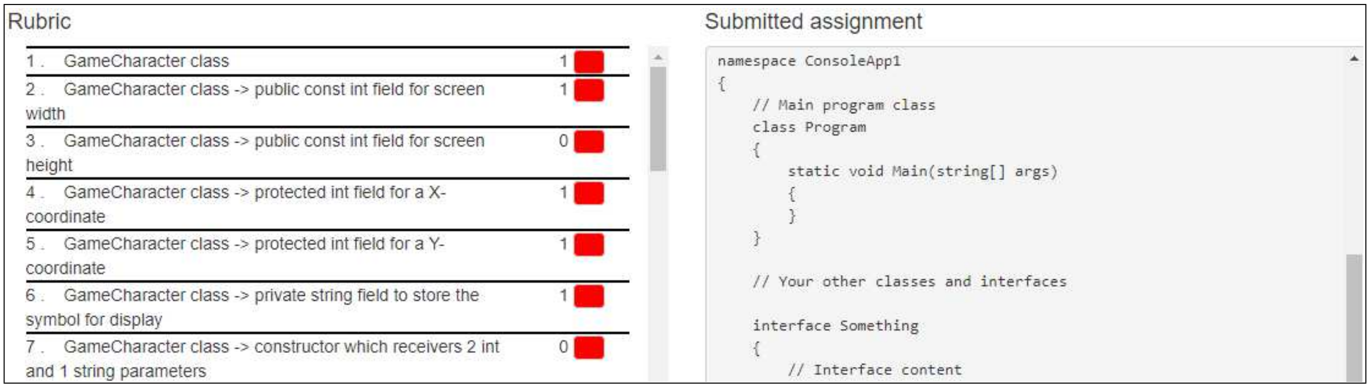
On average, marker 1 and 2 come very close to marking reliably. However, none of the marker pairs reach the acceptable level of agreement of 0.8 and the average, overall level of agreement between marker pairs is only 0.522. [12, p. 241] suggests that even for drawing tentative conclusions, used in exploratory research, agreements are only acceptable if they are higher than 0.667. All data below this minimum value should be rejected as unreliable. Thus, when looking at the individual results of markers, it would seem that their results are largely in disagreement with other markers. Such a process cannot be deemed as reliable and would result in possibly inaccurate feedback provided to the students whose assignments have been marked.

Delving a bit deeper into the differences between markers; Table II presents an overview of the percentage of assignments on which there were differences on the final mark assigned to specific assignments between pairs of markers. This data indicates that there is only one instance in which 2 markers were in complete agreement and that was on assignment 3 between marker 2 and 3. This only takes into account the final allocated percentage and not how the markers marked the individual rubric items or allocated errors. These results demonstrate the discrepancies which might occur when using multiple markers; even when using the exact same rubric they rarely arrive at the same result. It is these discrepancies which highlight the benefit that consensus moderation might provide, by improving the reliability of results through the opinions of multiple markers.

The results can be drilled down further by delving into the individual rubric items on the assignments. Table III lists the percentage of differences across the individual rubric items between marker pairs across all submitted assignments, e.g. there were 92 submissions for assignment 1 and each assignment rubric for assignment 1 contained 37 markable items. This results in 3404 items on which a pair of markers need to reach agreement. The first column in Table III lists this number of individual rubric items for each assignment.

Table III shows that even though the markers might yield similar final scores on their marking, they might arrive at these final scores by scoring the rubrics very differently. A good example of this is assignment 3 marked by the marker pair of markers 2 and 3. Their final results were the same, showing no diferences on the final scores, but on the individual rubric assesments they had different interpretations on 43% of the items.

The discrepancies between the final results of the markers and the agreement on the individual rubric items might be explained by the fact that markers also have the ability to indicate errors in the assignment, e.g. incorrect code syntax,



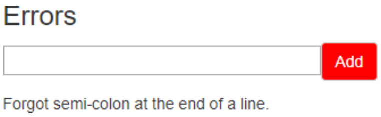
1 . GameCharacter class 1 🟥
2 . GameCharacter class -> public const int field for screen width 1 🟥
3 . GameCharacter class -> public const int field for screen height 0 🟥
4 . GameCharacter class -> protected int field for a X-coordinate 1 🟥
5 . GameCharacter class -> protected int field for a Y-coordinate 1 🟥
6 . GameCharacter class -> private string field to store the symbol for display 1 🟥
7 . GameCharacter class -> constructor which receivers 2 int and 1 string parameters 0 🟥

```
namespace ConsoleApp1
{
    // Main program class
    class Program
    {
        static void Main(string[] args)
        {
        }
    }

    // Your other classes and interfaces

    interface Something
    {
        // Interface content
```

Fig. 3. The Rubric Section of the Marking Interface



Forgot semi-colon at the end of a line.

Fig. 4. The Errors Section of the Marking Interface

TABLE IV
NUMBER OF ASSIGNMENTS DEEMED AS CONTAINING ERRORS BY EACH MARKER ON INDIVIDUAL ASSIGNMENTS

Assignment	Marker 1 and 2	Marker 1 and 3	Marker 2 and 3
Assignment 1	3	5	22
Assignment 2	7	7	22
Assignment 3	6	8	9
Assignment 4	3	16	26

which would result in automatic mark deductions. Table IV shows how many assignments each marker indicated as containing errors.

Similarly to the overall assignment marks, it can be seen in Table IV that the individual markers have very different opinions on what should be constituted as errors or apply different levels of strictness to their marking. Table V shows how the number of assignments with errors change if assignments are marked in pairs. When combining the results the number of errors are arrived at by selecting the number of errors presented by the marker which indicated the least amount of errors. The reasoning behind this is that both markers would have at least arrived at that minimum number of errors (according to count), regardless of what the actual errors were. This works as follows: if the first marker encountered 5 errors and the second encountered 3, it means both were in agreement that there are 3 errors (although one indicated 2 more). The number of errors on the assignment are then taken as 3.

Another aspect in determining marking consistency between markers is to determine how many assignments each marker deemed not to require marking, because it may be a copied assignment. In a programming subject this may be very subjective as what may seem to be copied code could simply be because of a variation on a provided example, strict adherence to the assignment specifications or requirements of the underlying programming technology. Thus, in attempting to identify plagiarised or copied assignments it may not yield consistent results when only relying on the opinion of a single marker. Table VI lists the number of individual assignments identified as copied by each marker for each assignment.

TABLE V
NUMBER OF ASSIGNMENTS DEEMED AS CONTAINING ERRORS BY MARKER PAIRS ON INDIVIDUAL ASSIGNMENTS

Assignment	Marker 1	Marker 2	Marker 3
Assignment 1	2	4	11
Assignment 2	7	11	4
Assignment 3	5	8	1
Assignment 4	2	2	8

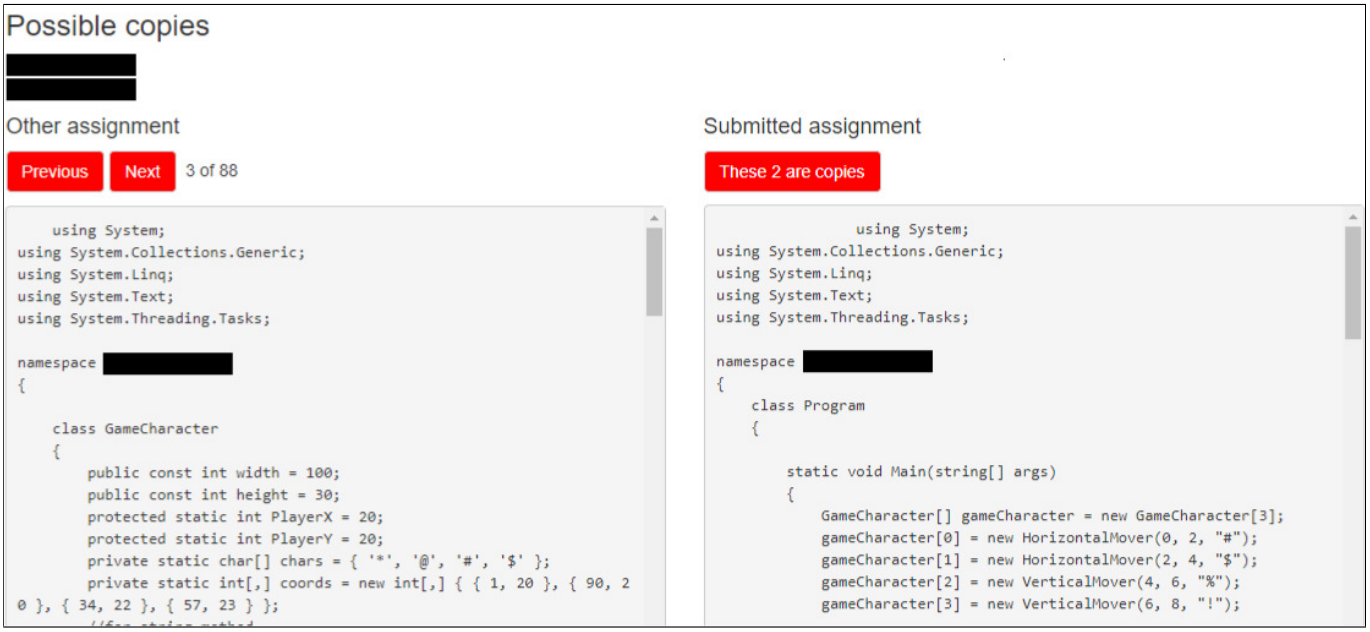
TABLE VI
NUMBER OF ASSIGNMENTS DEEMED AS COPIED BY EACH MARKER ON INDIVIDUAL ASSIGNMENTS

Assignment	Marker 1 and 2	Marker 1 and 3	Marker 2 and 3
Assignment 1	11	4	0
Assignment 2	1	12	4
Assignment 3	13	1	1
Assignment 4	15	4	0

Table VI highlights the difference in opinions which result when assignments are marked by different markers. Marker 1 is seemingly the most strict and indicates the most assignments as being possible copies, whereas marker 3 is the least strict. Table VII indicates how drastically the number of perceived copied assignments drop if only those assignments which are flagged as copies by pairs of markers (instead of individual markers) are taken as possible copies. The total number of possible copies for assignments 1 and 2 have dropped from possible maximums (according to Table VI) to 4 and 1, whereas assignment 3 and 4 have dropped to 0 possible copies each. This approach limits the number of assignments incorrectly flagged as copies and as such drastically limits the number of queries a lecturer would have otherwise received regarding assignments incorrectly flagged as copies.

V. DISCUSSION

From the results presented in Section IV it is clear that the results presented by individual markers vary widely. This



Possible copies

[redacted]

Other assignment

3 of 88

```
    using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace [redacted]
{

    class GameCharacter
    {
        public const int width = 100;
        public const int height = 30;
        protected static int PlayerX = 20;
        protected static int PlayerY = 20;
        private static char[] chars = { '*', '@', '#', '$' };
        private static int[,] coords = new int[,] { { 1, 20 }, { 90, 2
0 }, { 34, 22 }, { 57, 23 } };
```

Submitted assignment

```
                using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace [redacted]
{
    class Program
    {

        static void Main(string[] args)
        {
            GameCharacter[] gameCharacter = new GameCharacter[3];
            gameCharacter[0] = new HorizontalMover(0, 2, "#");
            gameCharacter[1] = new HorizontalMover(2, 4, "$");
            gameCharacter[2] = new VerticalMover(4, 6, "%");
            gameCharacter[3] = new VerticalMover(6, 8, "!");
```

Fig. 5. The Copies Section of the Marking Interface

TABLE VII
NUMBER OF ASSIGNMENTS DEEMED AS COPIED BY MARKER PAIRS ON INDIVIDUAL ASSIGNMENTS

Assignment	**Marker 1**	**Marker 2**	**Marker 3**
Assignment 1	4	0	0
Assignment 2	0	0	1
Assignment 3	0	0	0
Assignment 4	0	0	0

makes it very hard to recommend using multiple, independent markers to mark student assignments. There is a possibility that a student could either pass or fail an assignment based on the marker assigned to mark his or her assignment. To that end it might be safer to have multiple markers mark the same assignment.

The results presented in Section IV indicate that by requiring the opinion of multiple markers to determine whether an assignment is copied or contains errors, might drop the number of false reports. What is yet to be determined is whether this approach could be applied to the individual rubric items of an assignment. What this would entail is that the majority of markers need to agree that a rubric item is correct in order for the result to be taken as correct. To see what kind of effect this has on the final results of assignment marks, Figure 6 – 9 illustrates the mark distribution of individual markers versus the mark distribution of the majority consensus vote. These mark distributions do not include marks deducted for errors or copied assignments.

Figure 6 – 9 each contain a graph which shows the actual mark percentage distribution accross the various mark brackets, such as 0 – 10%, 11 – 20% up to 91 – 100% for each marker. The figures also each contain a graph which illustrates the trendlines for the same data. From the data presented in each diagram, it seems as if the *Combined* marking presents

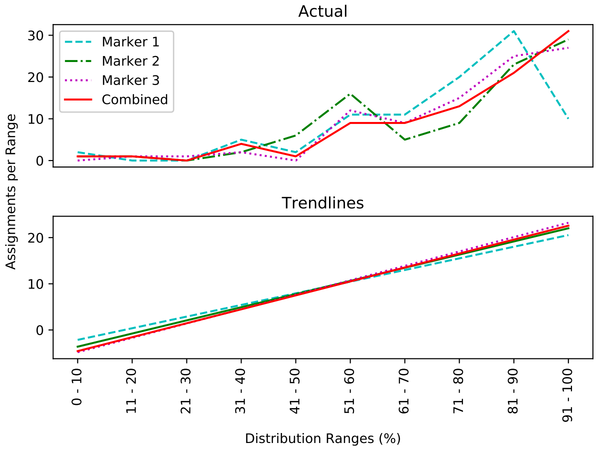


Fig. 6. Mark Distribution for Assignment 1

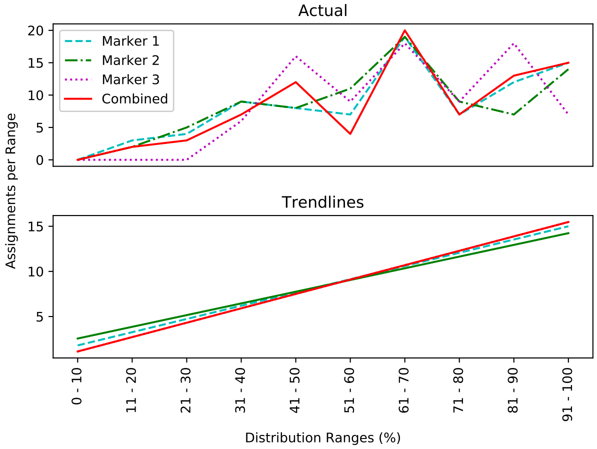


Fig. 7. Mark Distribution for Assignment 2

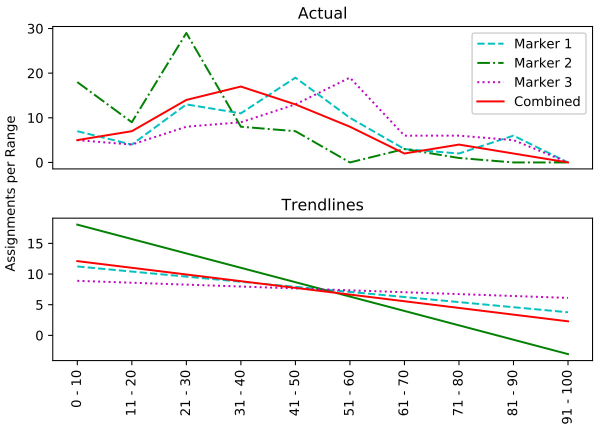


Fig. 8. Mark Distribution for Assignment 3

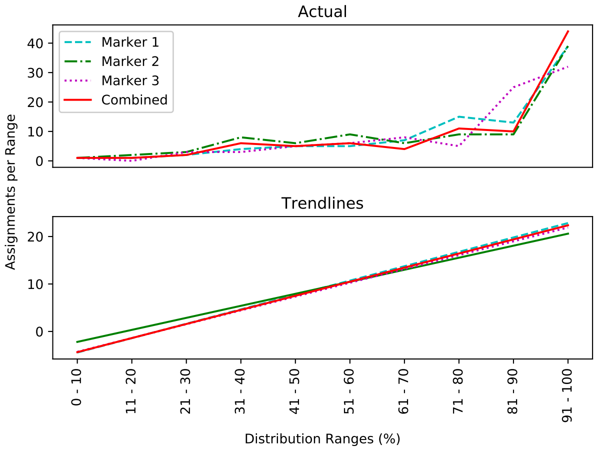


Fig. 9. Mark Distribution for Assignment 4

a more smoothed result. This is very apparent when looking at spikes in the graphs in Figures 6 and 8.

Figure 8 presents the results from Assignment 3, which seems to have been the most difficult if the marks are taken as the indicative factor. When the trendlines for this Figure is viewed, it is also clear that the *Combined* approach creates a trendline much less steep than that of the strictest marker, but also not as shallow as that of the least strict marker. If the trendlines for the other figures as scrutinized it also seems to indicate that the *Combined* approach tends to decrease the number of low assignment marks and boosts the number of top-end performers.

Table I indicated that there is a very low level of agreement between individual markers. To determine whether automated consensus moderation provides a higher level of reliability when compared to the marking of individual markers, Krippendorf's α was calculated for the individual rubric items of each marker per assignment, compared to the consensus marking results on the same assignments. These results are shown in Table VIII.

The results in Table VIII indicate that on average the individual markers have a reliability / agreement value of 0.811 towards the consensus moderation result. When compared with the average of 0.522 the 3 individual markers had among themselves; it would seem that consensus moderation provides a substantial increase in reliability. Another promising result is that the averages for each marker across all assignments are all above .7 and the average for all markers for each assignment also approaches .8; all of which indicates a level of inter-rater reliability which is acceptable.

TABLE VIII
INTER-RATER RELIABILITY BETWEEN MARKERS AND THE COMBINED APPROACH CALCULATED AS KRIPPENDORF'S α

Assignment	**Marker 1**	**Marker 2**	**Marker 3**	**Average**
Assignment 1	0.639	0.821	0.932	0.797
Assignment 2	0.855	0.855	0.835	0.848
Assignment 3	0.951	0.453	0.905	0.770
Assignment 4	0.829	0.748	0.903	0.827
Averages	0.819	0.719	0.894	0.811

TABLE IX
AVERAGE PERCENTAGE DIFFERENCE BETWEEN MARKER PAIRS ON INDIVIDUAL ASSIGNMENT RESULTS

Assignment	**Marker 1 and 2**	**Marker 1 and 3**	**Marker 2 and 3**	**Average Difference**
Assignment 1	16.22	21.62	18.92	18.92
Assignment 2	9.62	15.38	15.38	13.46
Assignment 3	25.81	17.74	37.10	26.88
Assignment 4	32.00	24.00	40.00	32.00
Average				22.82

Another way in which the results from the 3 markers and the *Combined* approach may be compared is by determining the percentage of rubric items on average a marker pair differ on per assignment. These values were then used to determine an average percentage of rubric item differences per assignment. This value is representative of how closely related the marking is of different marker pairs. Lower averages (approaching 0) are better. These results are shown in Table IX. As can be seen from the results there are relatively large differences of opinion between the various markers. Both assignment 3 and 4 have a larger than 25% average difference in marking between marker pairs. This illustrates how different (and possibly unfair) the final assignment results might be if different markers mark different assignments in the same group of assignments. When taken across all assignments, the 3 markers had a difference on opinion on 22.82% of the rubric items, which means excluding errors and assignments marked as possible copies, there could be a 22.82% difference between a student's results depending on who marked their assignment.

Table X lists how each individual marker's results compares with the consensus-based approach (i.e. the consensus-based approach is presented as a fourth marker). By scrutinising the average differences, the results indicate that the average difference between the individual markers and the combined approach is lower on all assignments than when comparing the results of individual markers with one another. The overall average difference across all assignments is 14.39%, which is 8.43% lower than the the average difference encountered when comparing the results from individual markers directly to one another. This suggests that individual markers would be closer

TABLE X
AVERAGE PERCENTAGE DIFFERENCE BETWEEN MARKERS AND THE COMBINED APPROACH ON INDIVIDUAL ASSIGNMENT RESULTS

Assignment	Marker 1	Marker 2	Marker 3	Average Difference
Assignment 1	13.51	8.11	13.51	11.71
Assignment 2	7.69	9.62	13.46	10.26
Assignment 3	8.06	19.35	19.35	15.59
Assignment 4	12.00	24.00	24.00	20.00
Average				14.39

to agreeing with a consensus-based approach than they would with each other's individual results.

VI. CONCLUSION AND FUTURE WORK

This study set out to determine whether the application of automated consensus moderation in a multi-marker environment is able to provide any benefit. From the results presented in this study, it is apparent that individual marker opinions might lead to widely varying results; even when using a preset marking rubric. This was apparent from the low level of inter-rater reliability between individual markers.

An automated consensus moderation effort can be generated by applying a majority vote rule to individual rubric items, errors and copies reported by individual markers. This does require some setup, such as a specialised on-line rubric. However, the individual markers are in greater agreement with the results of the combined marking than they are with each other individually. This suggests that the combined effort yields a result which more closely conforms to a normalised result; and as such is more reliable. The reason for this is that this type of consensus moderation excludes outlier results. Because multiple votes are required to present the final result, the student whose assignment is marked in such a fashion can be more confident that they are receiving an accurate and representative result. This also has the knock on effect of increasing the consistency across the marking of all assignments, since individual marker opinions are always consolidated into a final representative opinion.

The downside to this approach is that an assignment would always need to be marked by at least 2 markers, e.g. if 2 markers mark then majority vote could be enforced by requiring agreement between both markers. For 3 or more markers, the majority vote would always apply, e.g. for 3 markers a majority of 2 is required and for 4 markers a majority of 3 is required. When only 2 markers are employed, both would have to mark all assignments, for 3 markers each would have to mark two thirds and for four markers each would have to mark three quarters of the assignments. An alternative approach would be to always have just enough markers mark an assignment to ensure that a majority vote can take place, e.g. if there are 4 markers every assignment has to be marked by 3 markers. This will allow for a 2 out of 3 majority to be calculated instead of a 3 out of 4 majority.

Even though consensus moderation potentially requires more assignments to be marked, which might have a costly knock-on effect if markers are paid by assignment or per hour; it does provide a few benefits. One such benefit is an increase in marking reliability. A completed rubric may be seen as a form of feedback given to students. Reliable and accurate feedback on assignments (and tests), increase the chances of enforcing the knowledge retention of correctly answered questions and reducing the chance of knowledge retention for incorrectly answered questions [13]. If students do not realise their assignment has not been marked correctly, they may inadvertently assume incorrect information to be correct; which may be detrimental in further examinations on the same content. The negative effects of testing tend to persist over time, making it crucial to dispel possible incorrect knowledge retention as soon as possible [14]. Reliable marking serves to avoid this form of incorrect knowledge retention.

Other benefits of consensus moderation is that there are fewer assignments to mark per marker if more than 2 markers are used and, in cases where there are more than 2 markers, a possible faster turnaround on the marking process, because individual markers need to mark fewer assignments. In the end, these factors have to be weighed up for their possible advantages and disadvantages to determine if such an automated consensus moderation approach would yield a benefit to a selected subject or module.

Future work to be done in this study is to determine students' opinions as to the perceived differences in their results when comparing individual marking to consensus moderation and the benefits or drawbacks it might present. The marking of individual markers and the consensus result will also be compared with the opinion of expert markers to determine how close the various efforts are to the marking done by an expert. Achieving consensus on the number of errors per assignment was arrived at by selecting the lowest number of errors on which the markers were in agreement. A future study will be conducted in order to determine whether this is the most accurate reflection of the number of errors or whether another process might yield a result more consistent with that of an expert human marker.

REFERENCES

[1] D. R. Sadler, "Assuring academic achievement standards: from moderation to calibration," *Assessment in Education: Principles, Policy & Practice*, vol. 20, no. 1, pp. 5–19, 2013.

[2] V. Crisp, "The judgement processes involved in the moderation of teacher-assessed projects," *Oxford Review of Education*, vol. 43, no. 1, pp. 19–37, 2017. [Online]. Available: https://doi.org/10.1080/03054985.2016.1232245

[3] R. L. Lim, "Linking results of distinct assessments," *Applied Measurement in Education*, vol. 6, no. 1, pp. 83–102, 1993.

[4] S. Connolly, V. Klenowski, and C. M. Wyatt-Smith, "Moderation and consistency of teacher judgement: teachers' views," *British Educational Research Journal*, vol. 38, no. 4, pp. 593–614, 2012.

[5] A. Quinlan, *A Complete Guide to Rubrics: Assessment Made Easy for Teachers of K-college*. Rowman & Littlefield Education, 2012. [Online]. Available: https://books.google.co.za/books?id=IrIsG-hzN7cC

[6] E. J. Riddle and M. Smith, "Developing and using rubrics in quantitative business courses," *The Coastal Business Journal*, vol. 7, no. 1, pp. 82–95, 2008.

[7] B. Walvoord, D. Stevens, and A. Levi, *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Stylus Publishing, 2013. [Online]. Available: https://books.google.co.za/books?id=bDKFAwAAQBAJ

[8] B. M. Moskal and J. A. Leydens, "Scoring rubric development: Validity and reliability," *Practical assessment, research & evaluation*, vol. 7, no. 10, pp. 71–81, 2000.

[9] M. J. Warrens, “Cohen’s kappa is a weighted average,” *Statistical Methodology*, vol. 8, no. 6, pp. 473 – 484, 2011.

[10] M.-Y. Kan, J. L. Klavans, and K. R. McKeown, “Linear segmentation and segment significance,” in *Proceedings of the 6th International Workshop of Very Large Corpora*, 1998, pp. 197–205.

[11] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, “Normalization of non-standard words,” *Computer Speech & Language*, vol. 15, no. 3, pp. 287–333, 2001.

[12] K. Krippendorff, *Content Analysis: An Introduction to Its Methodology*, 2nd ed. SAGE Publications, Inc., 2004.

[13] A. C. Butler and H. L. Roediger, “Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing,” *Memory & Cognition*, vol. 36, no. 3, pp. 604–616, 2008.

[14] L. K. Fazio, P. K. Agarwal, E. J. Marsh, and H. L. Roediger, “Memorial consequences of multiple-choice testing on immediate and delayed tests,” *Memory & Cognition*, vol. 38, no. 4, pp. 407–418, 2010.