Evaluation of Humans and Software for Grading in an Engineering 3D CAD Course

Abstract

In skill-building courses such as an introductory 3D CAD course, instructors typically provide many assignments for students to practice and improve their 3D modeling skills. Frequent and accurate assessments give students the opportunity to identify errors and address deficiencies more efficiently, promoting quicker acquisition of the skill. In an ideal learning environment, students would be provided feedback at every class meeting, but that can be a daunting task as grading 3D CAD homework is difficult and time-consuming. The objective of this work is to compare human and software grading of student's 3D CAD files and quantify the speed, accuracy, and effectiveness each. A statistical analysis was performed on 5200 models from three different assignments to compare the two modes of grading. Better understanding of the different grading practices enables resource allocation based on strengths of humans and computers; resulting in a more efficient combination of resources. The results show that Graderworks© software (GW) was more accurate and repeatable than human graders (TAs) in quantitative comparisons: evaluating material, volume, shape, and sketches. TAs often made a few clerical errors per assignment that limited the effectiveness of the file management structure and subsequent calculations from manually entered fields like the name or username. However, a single change in the learning management software naming convention of files lead to a large scale clerical error with similar frustrations with automation of grading. Still, one of the biggest challenges we have experienced with human grading is the high variability in speed and accuracy of graders; an ANOVA test showed that error rates differ between TAs at a statistically significant level. TAs are effective at providing informative feedback that provides direction for improving the model, but it is a time consuming process. At this time, the software is not able to offer substantial and specific feedback to the students on how to improve, and it is recommended to use the computational grading tools in conjunction with human graders. Using the software to prioritize which files need TA feedback, those with similarity scores below a threshold value, may lead to a more efficient and effective use of resources to provide a quality feedback loop.

Introduction and Motivation

At Clemson University, the introductory engineering graphics course, ENGR 2080, uses SOLIDWORKS [1] to teach 3D modeling of parts and assemblies. To help the student improve their 3D modeling skills, students complete bi-weekly labs, weekly homework assignments, three projects, and take the Certified SOLIDWORKS Associates (CSWA) exam. The bi-weekly labs require students to produce a 3D CAD model of a part or assembly based on an engineering drawing. Labs generally take students less than 30 minutes to complete. The weekly homework assignments require the completion of tutorials and/or modeling parts or assemblies from the textbook [2]. Weekly homework assignments consistent of 3-5 parts or assemblies that the student must create in the CAD software.

Although assigning extensive opportunities to model helps with mastering 3D CAD software, efficiently grading the work is challenging. In the spring 2018 semester, 450 students were enrolled in the ENGR 2080 class at Clemson University. For 450 students, this results in grading $(4)(450) \approx 1800$ part files a week when only considering the weekly homework assignments. Generally, 3D CAD files take more memory on the computer and therefore loading the file into the CAD program usually takes longer than opening a simple file such as a PDF. 3D CAD files are complex and are often based on 100s or 1000s of parameters which makes grading them challenging, especially since there are many possible ways to arrive at the correct final geometry. Because of these complexities, four graduate student teaching assistants (TAs) graded the assignments for the ENGR 2080 class so that the approximately 1800 3D CAD files could be graded in less than one week and students receive feedback before the next assignment is due. An example of a weekly homework assignment is shown in Figure 1.

Create

Create only the following **Part** files:

1. ***Exercise 4*** on page 109 of "Beginners Guide to SOLIDWORKS 2017 - Level I". *Do not create the two tap holes labelled as M6x1.0 Tapped Hole, Default depth (2x).* Save the part file named as **username_EX4**.
2. Create the ***Circular Plate*** attached to this assignment. Save the part file named as **username_Circular Plate**.
3. Create the ***Slotted Plate*** attached to this assignment. Save the part file named as **username_Slotted Plate**.

Submit

Submit your work as follows:

- Create a folder named "**username-A2**" where username is replaced by your username (e.g. Dr. Yasmin would create a folder named: **yasmin-A2**).
- Put all three part files in the folder.
- Create a zip file of the folder.
- Upload the zip file (*.zip) into Canvas.

Figure 1. Instructions for Weekly Homework #2.

Assignment	Name	Points Possible	Points Lost	Late penalty	Grade
A2		100	0		100/100
	Description	**Possible**	**Lost**	**Comments**	
Formatting	**File Submission A2 (10)**	**10**			
	Files Named Correctly	3			
	Folder named correctly (username-A2)	3			
	Submitted .zip folder of parts	4			
Exercise 4	**Part Submission Exercise 4 (30)**	**30**			
	Linear pattern square indents (or equivalent)	5			
	Slot extrution (or equivalent)	5			
	Correctly dimensioned, fully defined	5			
	Units: mmgs	5			
	Assigned material (ABS)	5			
	Volume: *232908.7385 cubic millimeters* (235237.8259 - 230579.6511)	5			

Figure 2. Part of the rubric for Weekly Homework #2. See Appendix A for complete details.

TAs complete a rubric for each student for each assignment. Rubrics list the 3D model files that the student should have turned in along with several items to evaluate on each file including checking for specific features of the part, the correct units, the correct shape, the correct volume, the correct material, and checking if the sketches are fully defined (See Figure 2). To speed up the grading process and reduce subjectivity, each item is graded as correct or incorrect with no partial credit given. The TAs digitally fill out the rubric for each student and create a PDF of the completed rubric and return it to the students via the course management system gradebook. Key attributes of this system of grading are 1) the grader can quickly fill out the rubric without having to reflect on the requirements of the assignment, 2) subjectivity is limited by having standard grading items to evaluate, 3) feedback is returned to the student electronically, making it more accessible and ensuring the feedback cannot be lost, and 4) an electronic record of errors is collected for assessment purposes. The following questions still lingered: Do the graders correctly fill out the rubrics? Do all the graders have adequate and consistent accuracy of grades?

A trade-off exists between evaluating quality with an "ideal rubric" and what is practical to implement within given resources. One limitation of the week homework rubrics is that they do not attempt to measure design intent [3]; "strategies for incorporating maximum design flexibility and minimizing design failure" [4]. Several researchers have examined the learning objective outcomes for an engineering graphics class and then used this data to create rubrics for assignments that test the learning objectives [5]. Pedro et al. used five criteria to classify the quality of CAD models. These included 1.) Models are valid with no errors or warnings while opening, 2.) Models are complete and include all the design relevant information, 3.) Models are consistent and can be easily edited, 4.) Models are concise and do not have unneeded or duplicate information, and 5.) Models express design intent [5].

Computational Grading Tools in the Literature

Several researchers have sought to reduce the workload of grading engineering graphics homework using automated tools. Because of the challenges with grading 3D CAD files, some instructors have investigated assisting or replacing human grades with software.

Ault and Fraser made a digital CAD grading tool to evaluate student CAD files. Their system grades by assessing the number and types of features, checking if a dimension exists within a sketch, and evaluating the shape of the model by evaluating if the volume is correct [3]. A limitation of this method is that looking for a specific dimension might not be the best, since there are multiple ways to dimension a single object. For example, the student might choose a different datum. Also, the part files presented by Ault and Fraser are relatively simple and do not reflect the complexity of most 3D CAD files.

Gonzales—lluch et al. use a commercial tool called *Model Quality Tester,* which was designed primarily to enforce consistency in CAD models at a medium or large enterprise. Gonzales-LLuch et al. mapped CAD rubrics from the literature to the abilities of the software tool and found that in general, the tool is not sufficient for automated grading of CAD files [6].

Hekman and Gordon evaluated using an automated system to grade AutoCAD files that focused on 2D drawings [7]. In this method, students email homework solutions to a dedicated email address. The automated system then reads the emails and evaluates the 2D drawing(s) submitted. The advantage of this system is that students could receive quick feedback, although the grading method's generalization to 3D CAD is not apparent.

Ingale et al. developed a tool in Matlab which would evaluate section views submitted as picture files by the students. Their tool had the advantage of being CAD software independent since it only used image files and image recognition techniques. However, the submitted picture files had no dimensions shown, and the tool could not process other common views like the top, right,

and front. In this method, students email homework solutions to a dedicated email address. The automated system then reads the emails and evaluates the 2D drawing(s) submitted [8].
Renu et. al. investigated assigning partial credit using a computational shape similarity score. The computational tool was able to produce a better distribution of student scores than a simple pass or fail metric like comparing the mass of the student's file to the solution's mass [9]. The technique used by Renu et. al. using a point cloud based geometric similarity comparison algorithm is similar to the work Paravati et. al. which evaluated 3D animations using point clouds [10]. Paravati's technique sought to train an automated grading algorithm and then compare the algorithm with manual grading by the course instructors.
Kirstukas developed a tool for automated grading of students files in NX CAD [11]. The tool checks for common student mistakes such as incorrect units, undefined sketches, banned sketch constraints, dimensions, correct shape, and more. The correct shape is evaluated based surface area and volume of the student's file compared to the correct answer "gold standard" file given to the tool by the instructor. A limitation of Kirstukas tools is that it is not clear how to check if the dimensions are correct since there exist many ways to draw the same object. A strategy for drawing an object might require multiple features with simple sketches while another strategy would use less features with more complicated sketches. The tool would need to determine that both strategies result in the same final shape with the same overall dimensions even though they were drawn using different strategies. Using 19 student models, Kirstukas compared manual grading to the automated tool grading and found a coefficient of determination value of 0.65 ($R^2 = 0.65$) [11].

Computational Grading Tools in Certification Examinations
The Certified SolidWorks Associate (CSWA) exam from SOLIDWORKS [12] is an auto-graded examination that tests users on basic knowledge of how to create parts and assemblies using SOLIDWORKS. Each question on the exam is pass/fail with no partial credit. The exam requires students to answer multiple choice questions about engineering drawings, model several parts from engineering drawings, and to assemble some device together correctly. The students are asked to report the mass of each part they model and the center-of-mass for the assemblies. Upon submission of the examination, the testing software instantly provides a numerical score for each question, the cumulative score, and whether the score was high enough to be considered passing. The mass and center-of-mass values inserted by the student must be within 1% of the correct answer value to be marked correct. The exam assumes that having the correct mass or center of mass means that the user has correctly drawn the shape, although it is not clear that this assumption is valid, but is more of a heuristic. Also, pass/fail grading can result in bi-modal distributions which typically are not desirable [9]. As of August 2018 approximately 210,000 SOLIDWORKS users have passed the exam which indicates its wide acceptance as a measure of basic 3D modeling skills [12].

Summary of methods
Table 1 summarizes the state of the art methods and research gaps. In summary, the literature makes a compelling case for using rubrics in grading and using automated grading. However, the literature lacks clarity on how automated grading compares to manual grading. A statistical study comparing automated grading of 3D models and manual grading would enable CAD educators to make informed decisions about when to use automated and manual grading.

Table 1 Summary of the literature

Author(s)/Developers	Method	Shortcoming(s)
Ault and Fraser [3]	Feature recognition, dimensioning, shape evaluation based on volume	Variability in dimensioning methods and shape assessment based on volume leads to incorrect grades being assigned.
Gonzales—lluch et al. [6]	Model Quality Tester	Designed for commercial use and not for evaluating student work.
Hekman and Gordon [7]	Automated system to grade 2D drawings	Not extensible to 3D solid models
Ingale et al. [8]	Assessment of section views	False positives are inherent
Renu et al. [9]	Geometric similarity algorithm assigns partial credit	Not clear if the partial credit scores assigned correlate with manually assigned scores
Kirstukas [11]	Software tool uses the CAD API to evaluate student's models using several metrics.	Not clear how the tool can evaluate models drawn using a different strategy. The sample size for comparison the automated tool to manual grading is small ($n = 19$).
SOLIDWORKS [12]	Assessment based on numerical values entered	"Pass/Fail" method limits feedback provided to students

Computational Grading Tool Used at Clemson University

To reduce the workload on the TAs, the bi-weekly labs are graded using Graderworks (GW), a SOLIDWORKS grading support program. GW loops over student's files, folders, or .zip files and analyzes SOLIDWORKS files. GW preforms the following operations:

1. extracts mass-properties from each file including mass, volume, density, surface area, and center of mass coordinates.
2. analyzes the shape of a student's model compared to the correct answer model using the D1 geometric similarity algorithm. Cardone et al. gives an overview of the D1 algorithm and other geometric similarity algorithms [13]. The applicability of the D1 algorithm to 3D CAD model grading was first investigated by Renu et. al. [9].
3. extracts feature information from each file and, if the file is a part file, obtains the constraint status of each sketch (fully defined, under defined, or not solvable) and determines who authored the files by parsing the file or folder name.

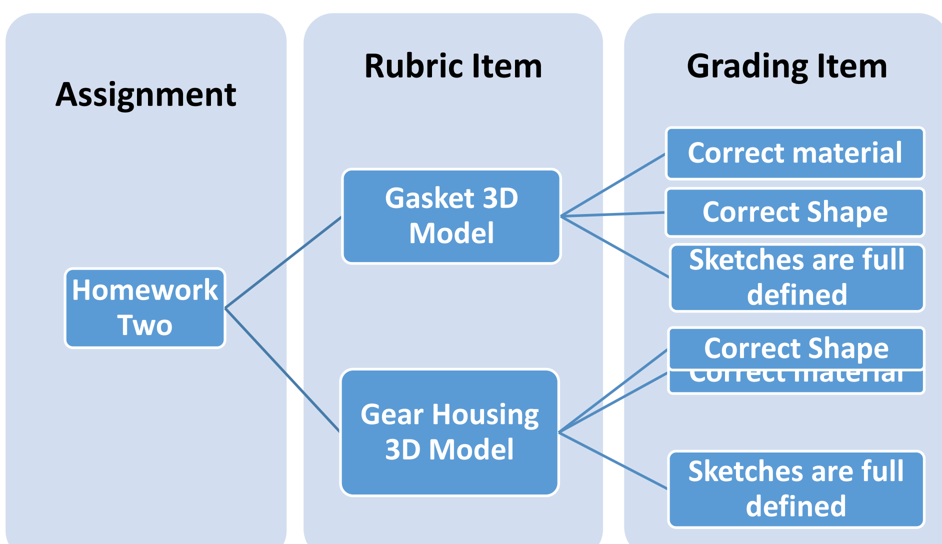


Figure 3. Show the hierarchy of assignments, rubric items, and grading items.

The software enables the instructor to generate rubrics for assignments using representative model files of a correctly designed model that reflects what students are expected to submit. Within each rubric, instructors can specify which features to grade: volume, material, composite shape score, center of mass, and check for fully defined sketches (see Appendix B for more detail). However, similar to the rubrics used by the graduate student graders, GW rubrics do not attempt to measure design intent in any manner. Although GW was used to grade the bi-weekly labs, it was not used to grade the weekly homework assignments because it was not certain GW could replicate the same level of feedback performed by the TAs.

Research Objective

In general, computational tools for grading 3D CAD files do have many advantages however it is not clear how they contrast to human graders. The objective of this study is to evaluate the strengths and weaknesses of both human and software grading of student's 3D CAD. The human and software grading practices are evaluated by comparing scores on grading items for three assignments which accounts for 9% of the course grade. Software and TA disagreement was investigated by one of the experienced faculty instructors of the class. Next, the TA graders were evaluated by comparing the number and types of errors that they made relative to each other. A description of the methods used to compare the human graders with the software graders is provided with results for three assignments.

Methods

To evaluate the strengths and weaknesses of both human and software grading of student's 3D CAD files, the rubric items scores from three homework assignments were compared (See Appendix A for complete TA rubrics used in this study). All four TAs each had prior experience grading for the course. The TAs completed rubrics for each student for weekly homework assignments. The individual scores on each grading item for each student was saved for academic assessment purposes and this study. The four graders had one week to complete grading each assignment.

GW was configured to have a similar rubric for each assignment, and all the student's files were processed and graded again, this time using the automated grading process (See Appendix B for complete rubrics used by GW for this study). The amount of time required to process the assignments using GW varies based on the number of files each student submits. For the assignments used in this study, the analysis time took between 90 minutes and 150 minutes. A Microsoft surface laptop with an i7 processor, solid state hard drive, and 8GB of memory analyzed the files. The software required approximately 3 seconds per file to extract data and perform its analysis. The speed of the software is dependent on the computer running the software and was not a significant consideration in this study. Assignment 4 requires adding more features and complexity to the same models students made for assignment 3. As a result, assignment 4 requires fewer files be submitted, but more features are analyzed (See Table 2). GW compares each file to rubric items and then computes a score for each rubric item and a comprehensive score for the assignment. Grading information is exported as an excel file which can be used to set up a mail-merge to email each student a completed rubric if desired. Each row of the output contains how a student performed on each grading item and list each sketch that is not fully defined.

Table 2. Assignment Metrics from GW

Assignment Number	Number of students who's files were analyzed by GW	Number of required files for assignment	Number of files analyzed	Number of features analyzed
2	433	3	1274	26254
3	441	5	2243	57251
4	433	4	1731	58226

Username	Volume Grade of Circular Plate score	Correct Material Grade of Circular Plate score	Shape Grade of Circular Plate score Partial credit applied	Fully Defined Sketches Grade ofCircular Plate score	List of Problem Sketches for Circular Plate	Volume Grade of Ex 4 score	Correct Material Grade of Ex 4 score	Shape Grade of Ex 4 score Partial credit applied	Fully Defined Sketches Grade ofEx 4 score	List of Problem Sketches for Ex 4	Volume Grade of Slottted Plate score	Correct Material Grade of Slottted Plate score	Shape Grade of Slottted Plate score Partial credit applied	Fully Defined Sketches Grade ofSlottted Plate score	List of Problem Sketches for Slottted Plate
student1	100	100	100	100		100	100	100	0	[Sketch1 is Under Defined]	100	100	100	100	
student2	100	100	96.45	0	[Sketch2 is Under Defined]						100	100	95.75	100	
student3	0	100	94.22	0	[Sketch2 is Under Defined]	0	100	81.88	0	[Sketch3 is Under Defined]	0	100	89.75	0	[Sketch2 is Under Defined]
student4	100	100	100	100		100	100	100	100		100	100	100	100	

Figure 4. Part of the GW output for homework assignment two.

In instances where the TAs and GW disagree, one of the instructors for the course manually spot checked the files to determine if GW correctly graded the files. No mistakes were found by GW, which was expected since GW is an algorithm. The purpose of this study was not to determine whether human graders made errors, but rather to understand what types of mistakes and how often. Many of the mistakes made by students were very small and in these instances, GW took off a few points from the shape score and the TAs did not identify the mistake. For example, some students rounded the depth of an extrude cut to the nearest whole number when they should specify the dimension to one decimal (72 mm vs. 72.2 mm). The TA graders were unable to identify these small mistakes while GW took off a few points from the student's composite shape score. Manually identifying the reason GW took off a few points proved to be a tedious task! Both sets of grading data, from the graduate student graders and GW, was combined using a SQL join statement on the student's email address. Then, the student's name or email address was replaced with a random number to anonymize the data. The resulting joined data table contains how GW scored each student on the rubric items and how the TA's scored each student on the rubric items in the TA's rubric. Not all the rubric items overlap between the TA's rubric and GW rubric. For example, GW does not check to see if the student put their university username in the file name of their folder, but the TA's do. While comparing the TA and GW data, it was found that on assignment two the TA's performed clerical mistakes on 2.4% of the student's rubrics. Clerical mistakes include:

- TA put the wrong username on the rubric.
- TA misspelled the student's username when filling out rubric.
- Student misspelled his/ her username or did not put the username in the .zip file's name. As a result, the TA also misspelled or did not include a username on the rubric.

Running GW also had some challenges. For example, Clemson University changed how suffixes (for example Jr. Sr. or III) of student's names are shown in the learning management system (LMS). The LMS renames student files and includes the student's actual name in the file name when submissions are batch downloaded from the LMS. GW parses the file or folder name of each downloaded file to determine which student the files belong to. Since Clemson University dropped the use of generation suffixes in the naming convention mid-semester, the change in name confused GW. For example, if a student were named John Brown Jr., the LMS renamed files as john_brown_jr_hw2.sldprt at the beginning of the semester, but later in the semester it changed to john_brown_hw2.sldprt. Although the problem was resolved, this short example highlights some of the complexities of using automated tools.

For this study, if the TA or GW was missing data for a grading item, the student's data was excluded from the calculations for that specific grading item. Only instances where both processes graded the same item were evaluated in this comparison. These instances fall into a few key categories such as checking the accuracy of features including: material, volume (to within 1% of the correct volume), shape, and fully defined sketches. For each grading item comparison, the total number of times that the TA and GW agreed or disagreed that a grading item should be marked correct or incorrect was recorded. For each grading item, a p-value is calculated where the null hypothesis is that the TAs and GW agree on how the items should be graded. The percent agreement rate was calculated for each grader for each grading item. The results are presented in Figure 5, 8, and 11. The percentage of students who received a passing score on the grading item is shown for the three assignments in Figures 6, 9, and 12.

Finally, the grade distribution as calculated by the TAs and GW for each assignment is shown. However since the rubrics do not fully align a direct comparison is not valid. The TA rubrics give ten points to students for turning in their work in the correct format (See Appendix A for complete rubrics). These items include submitting a .zip folder, putting their username in the .zip folder name, and giving the individual parts the proper name. The GW grading item point values were adjusted to best match the point values (as a fraction of the total points on the assignment) of the TA rubric. The following formula can be used to convert a student's final assignment grade on the TA rubric to the assignment grade on the GW rubric. This conversion assumes that the lowest score for the TA rubric is 10 point since formatting the assignment is not necessarily a measure of modeling skill but is allotted 10 points on the TA rubric. s_{GW} is the total score for the student calculated by GW. s_{TA} is the total score for the student calculated by the TAs.

$$\frac{s_{GW} - 0}{100} = \frac{s_{TA} - 10}{90} \tag{1}$$

The purpose of showing the grade distribution as a histogram is to compare the distribution of the TAs and GW to ensure their similarity. Figures 7, 10, and 13 show the histograms where the GW scores were converted to be in the same scale as the TA scores using Equation (1).

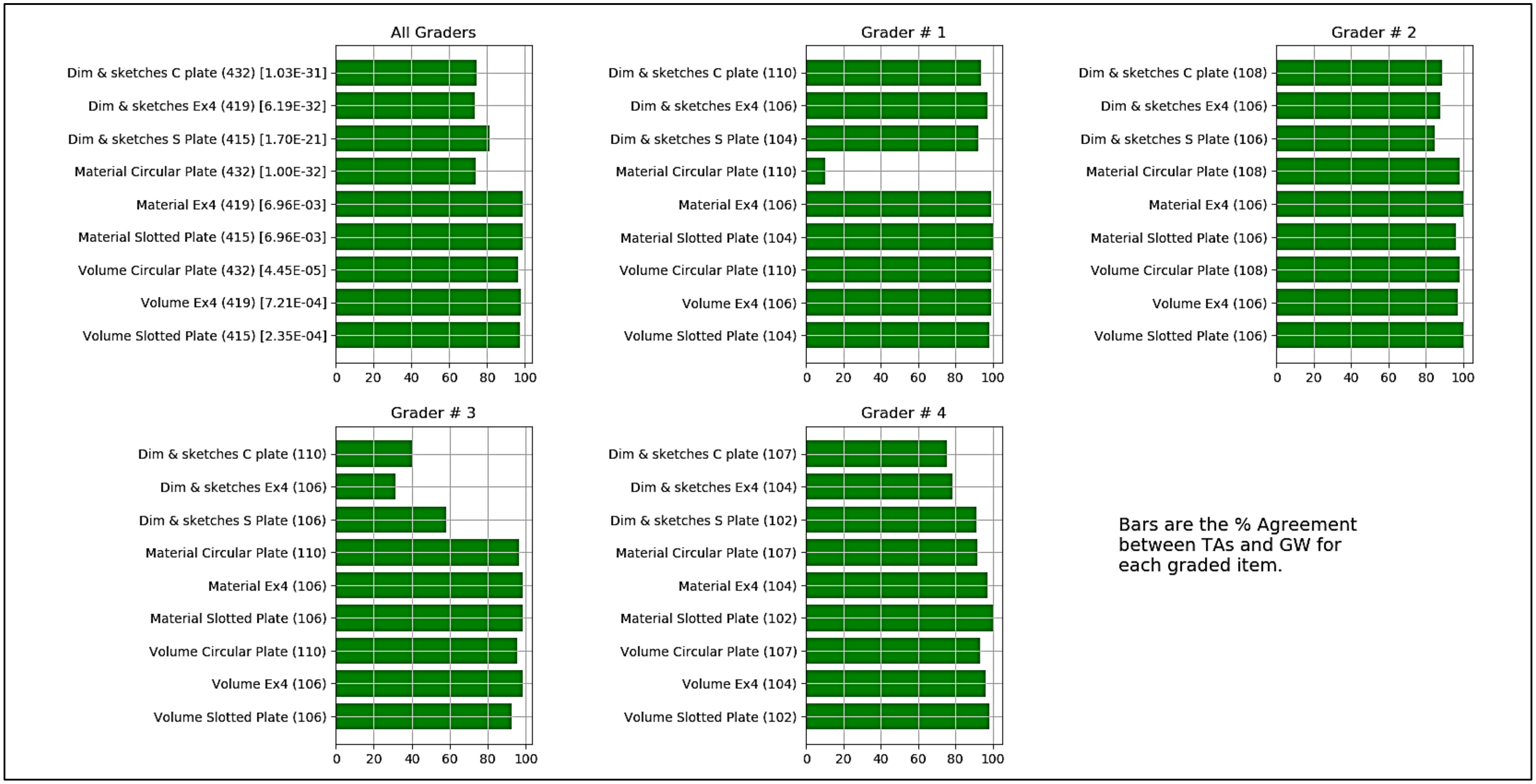


Figure 5. Results for Assignment 2. Each bar graph shows the percent agreement of the TA and GW for each grading item. The number in parenthesis is the number of comparisons performed. The number in square brackets is the p-value. C plate is Circular plate. S plate is Slotted plate. Dim is dimension.

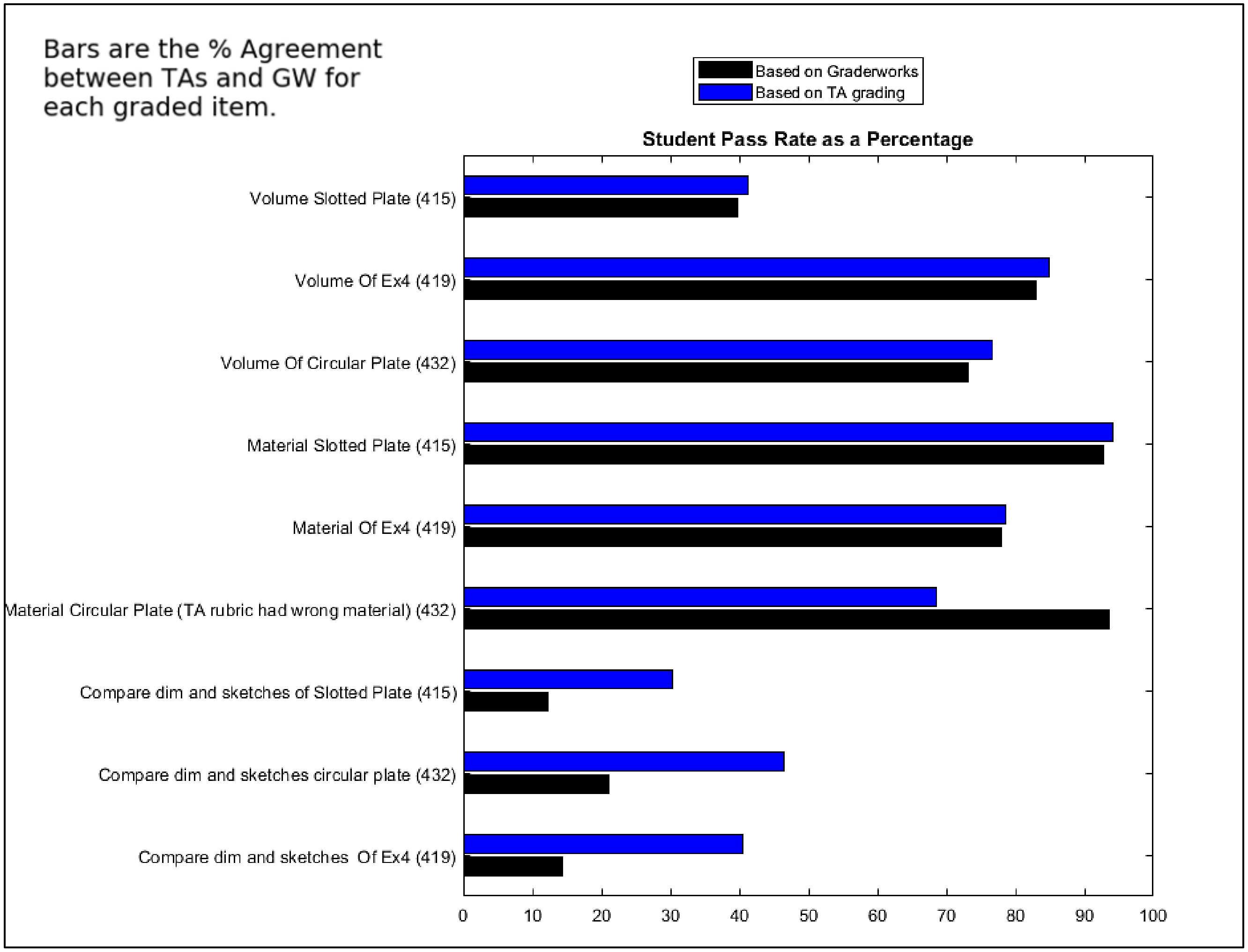


Figure 6 Student pass percentage for grading items in assignment 2. Student pass perentage is the percent of students who earned full credit for the particular grading item. The number in parenthesis is the number of students evaluated.

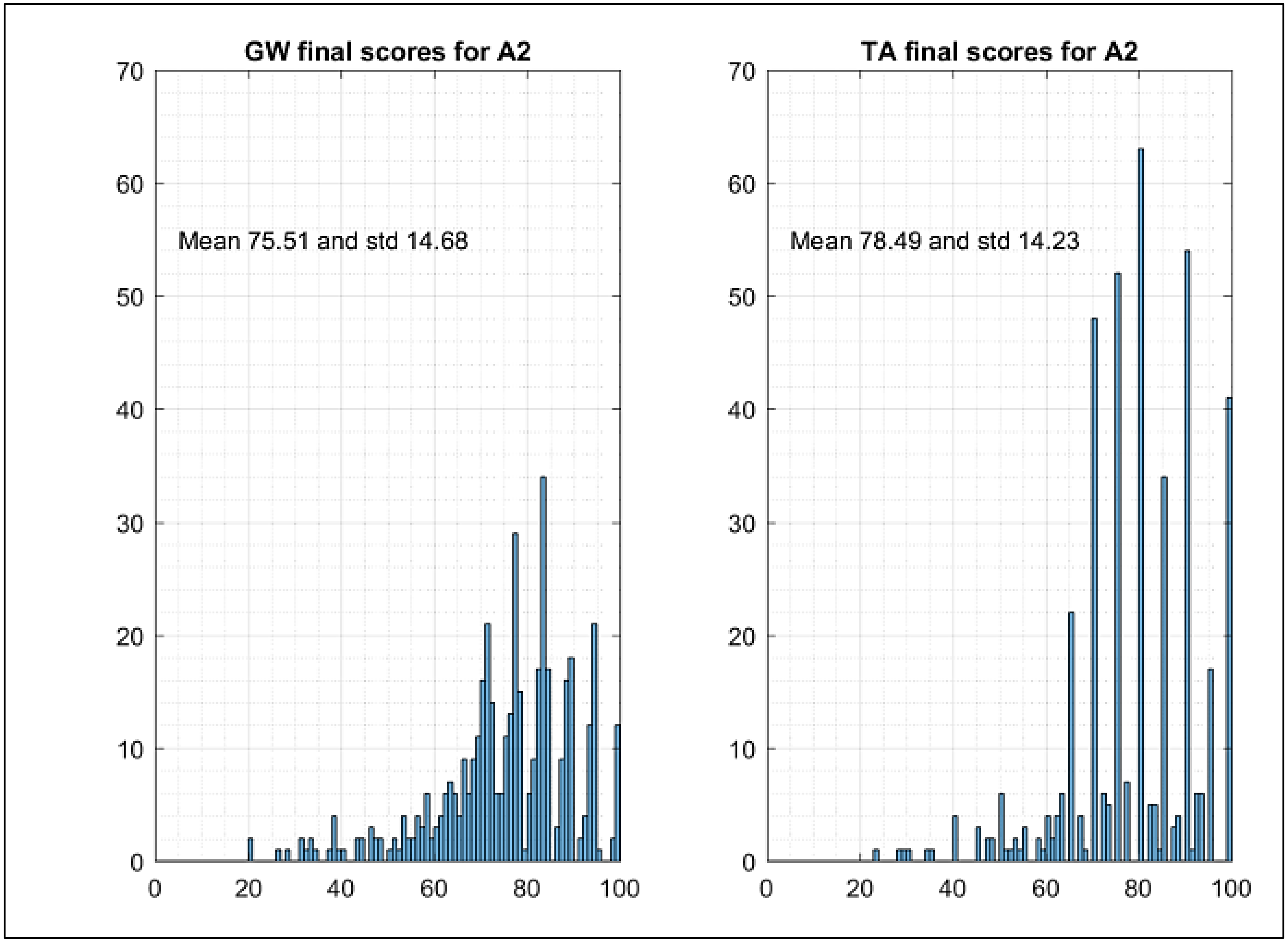


Figure 7. Histogram of assignment scores for GW and the TA comparing frequency (y-axis) and spread of grades (x-axis).

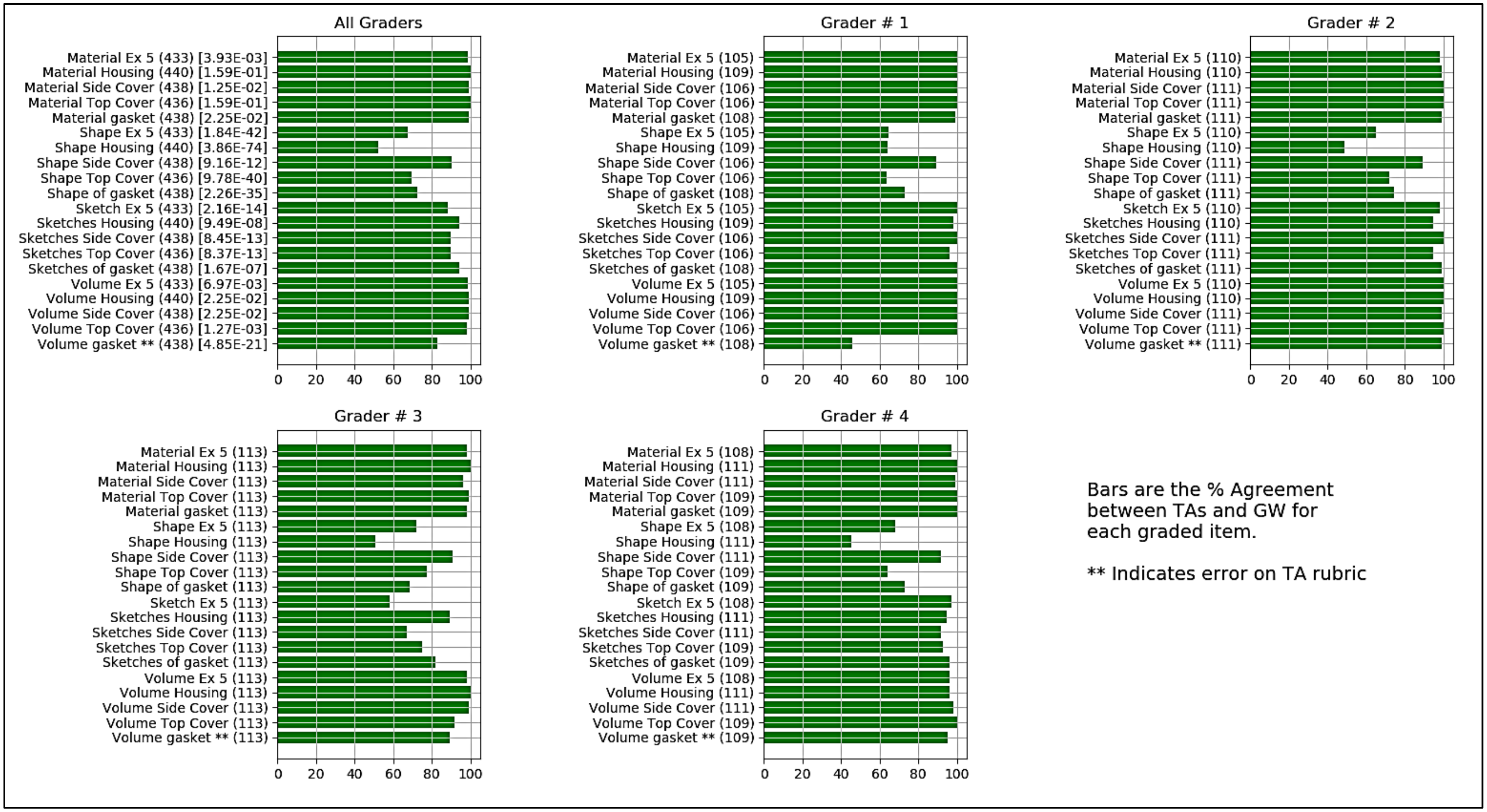


Figure 8 Results for assignment 3. Each bar graph shows the percent agreement of the TA and GW for each grading item. The number in parenthesis is the number of comparisons performed. The number in square brackets is the p-value.

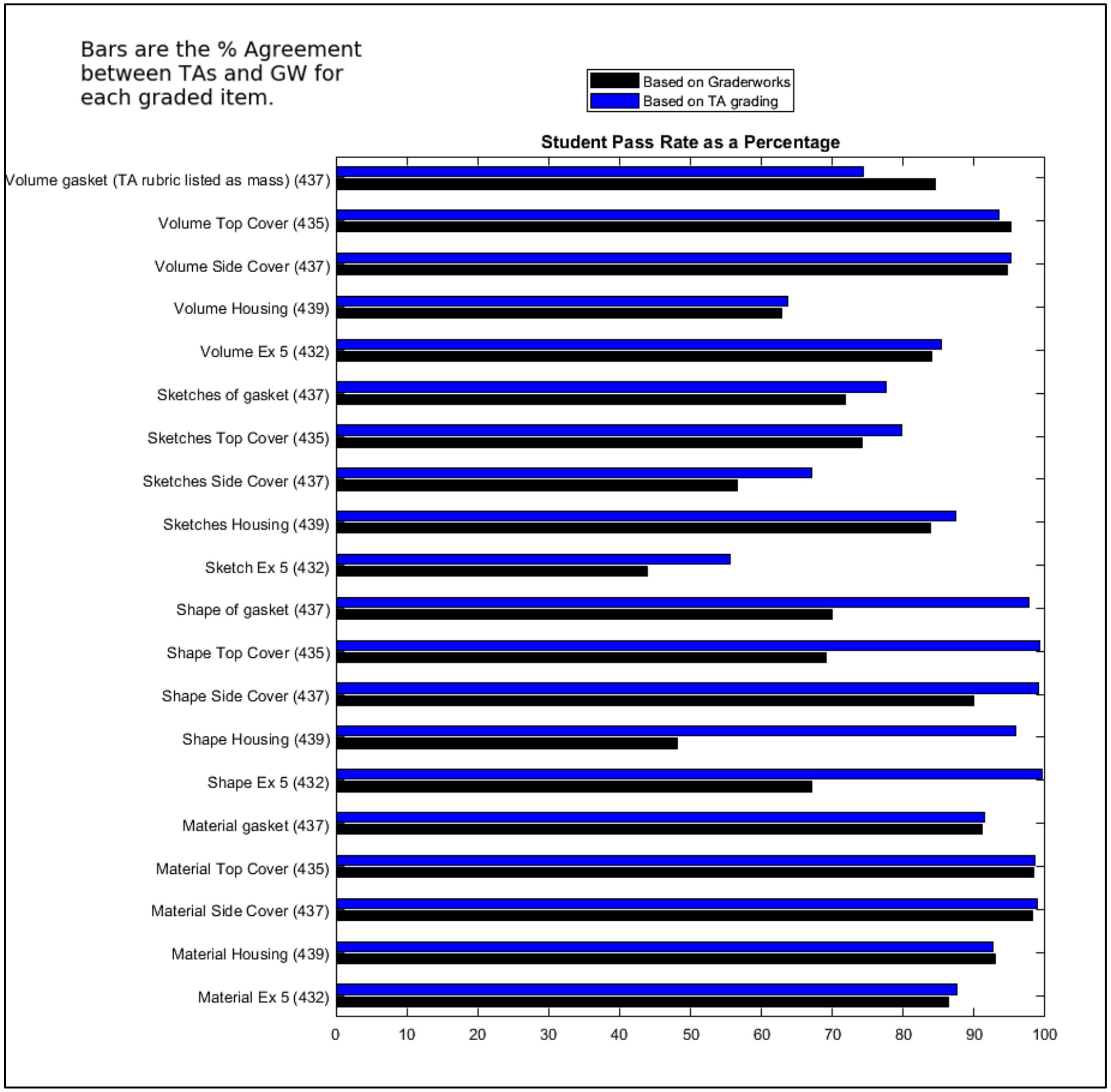


Figure 9 Student pass rate for grading items in assignment 3. Student pass perentage is the percent of students who earned full credit for the particular grading item. The number in parenthesis is the number of students evaluated.

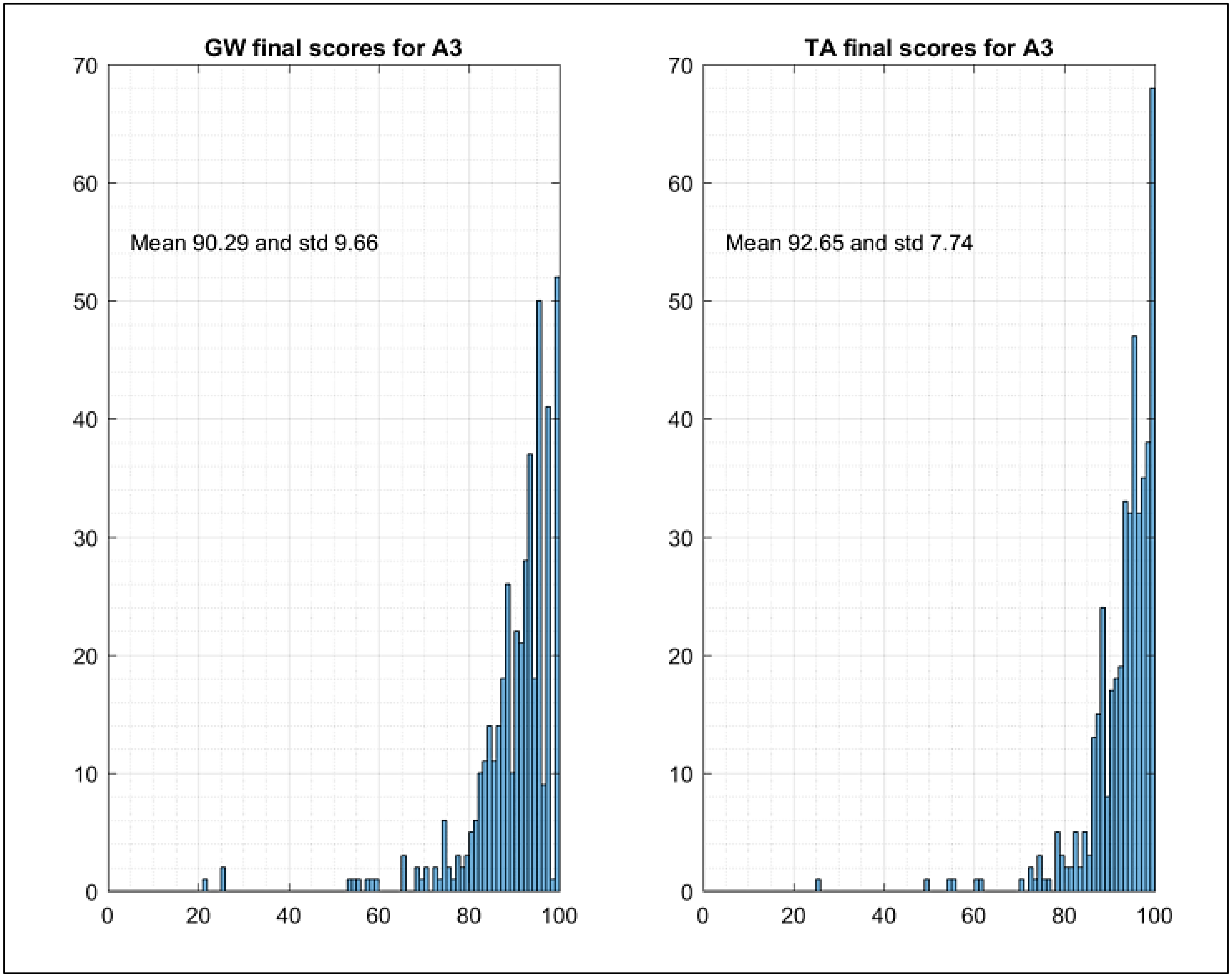


Figure 10. Histogram of assignment scores for GW and the TA comparing frequency (y-axis) and spread of grades (x-axis).

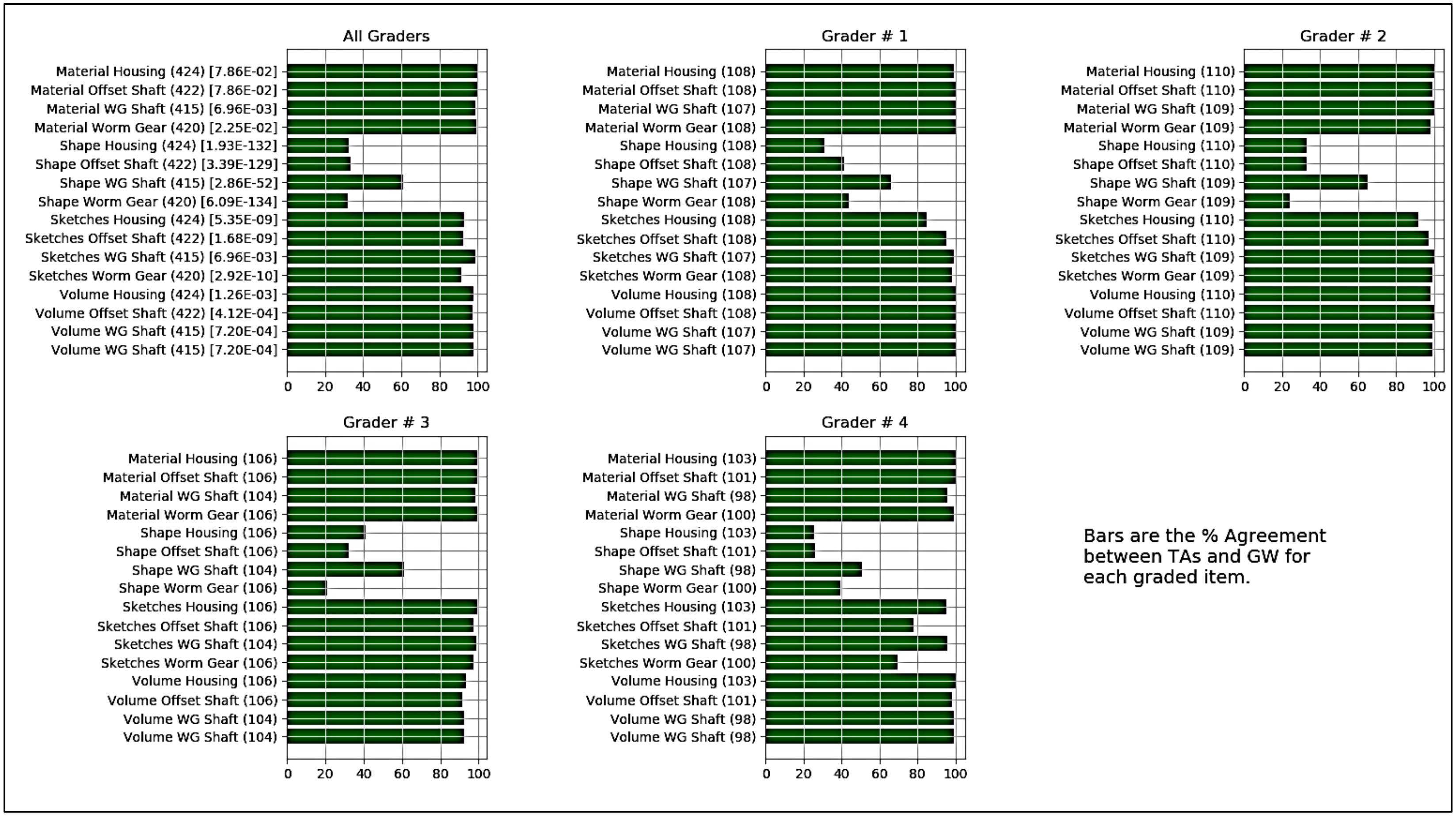


Figure 11 Results for Assignment 4. Each bar graph shows the percent agreement of the TA and GW for each grading item. The number in parenthesis is the number of comparisons performed. The number in square brackets is the p-value.

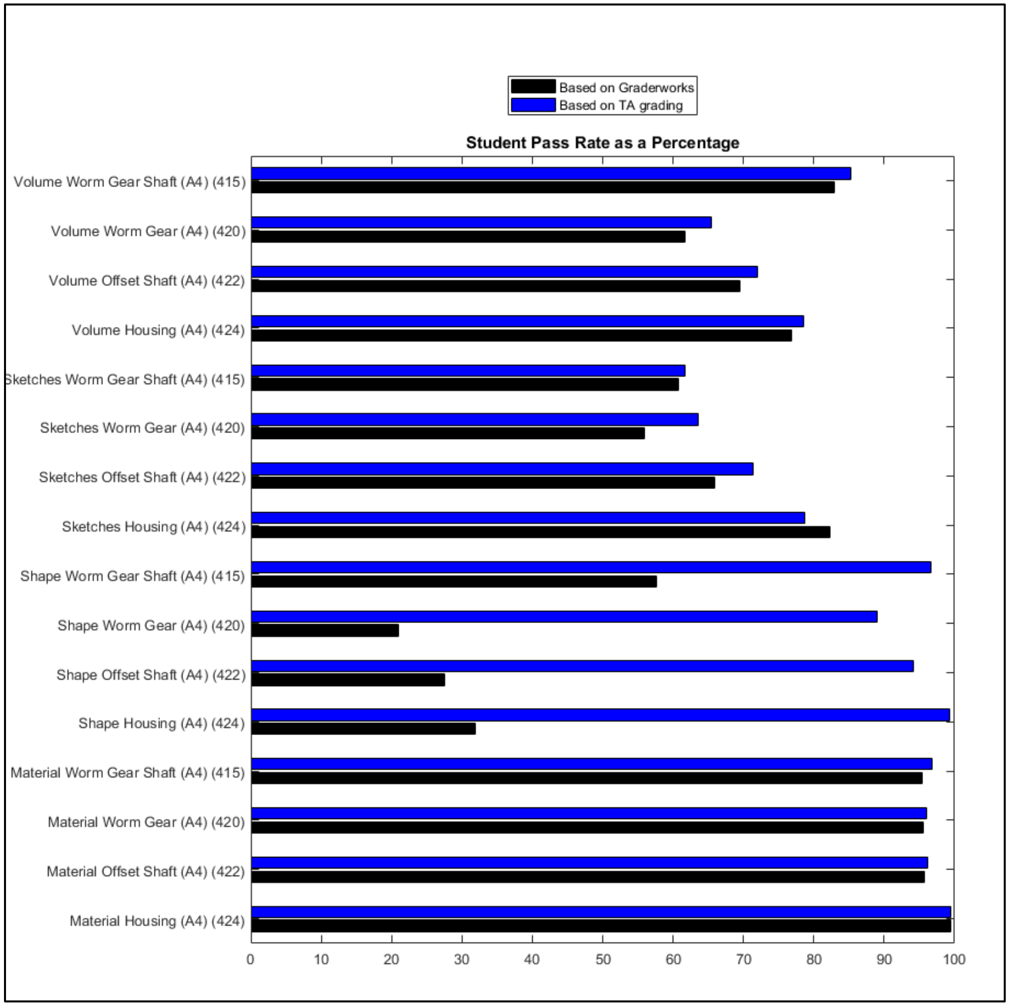


Figure 12. Student pass rate for grading items in assignment 4. Student pass perentage is the percent of students who earned full credit for the particular grading item. The number in parenthesis is the number of students evaluated.

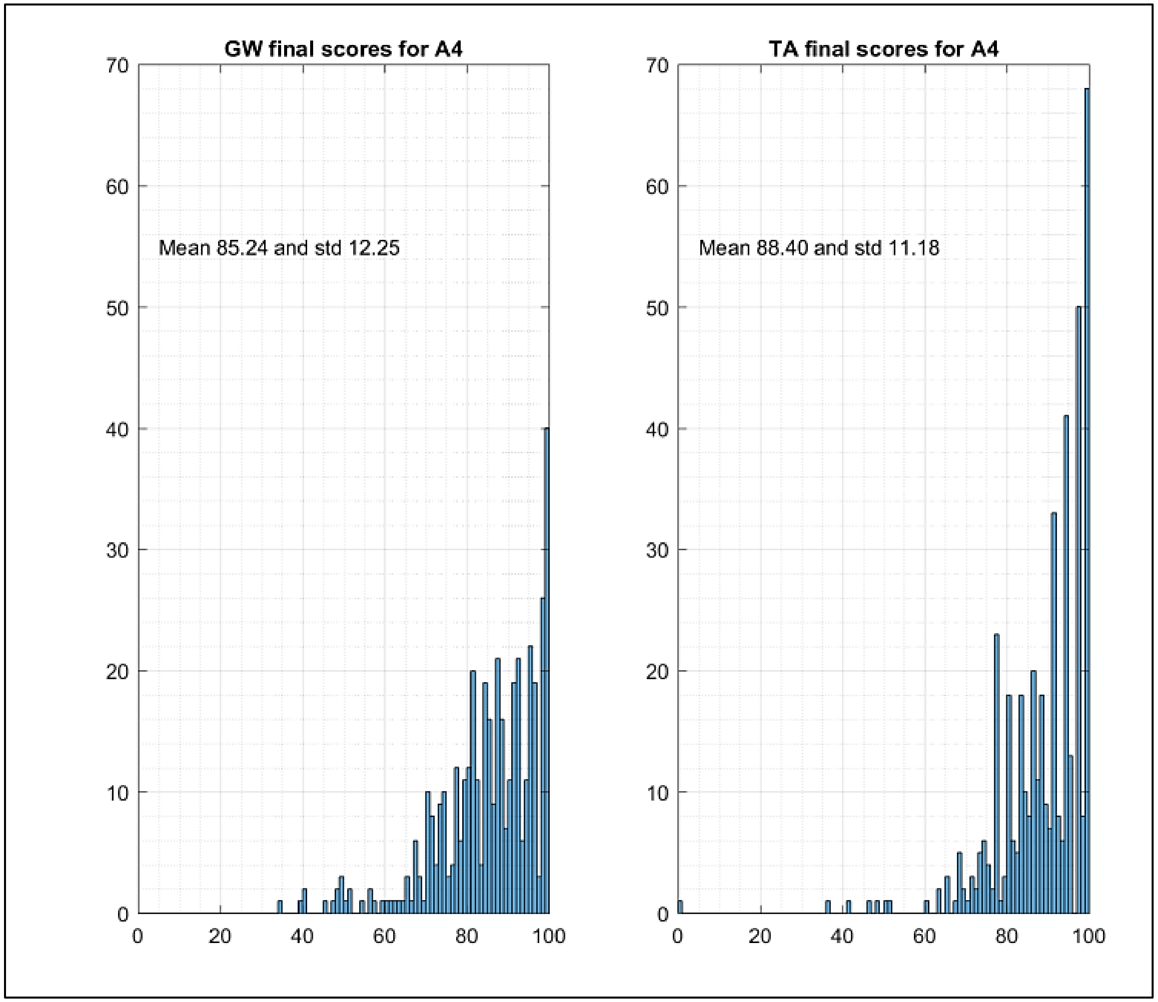


Figure 13. Histogram of assignment scores for GW and the TA comparing frequency (y-axis) and spread of grades (x-axis).

Results

Of the 45 grading items compared only 4 had a p-value higher than 0.05. For the other 41 grading items with a p-value smaller than 0.05, the null hypothesis is rejected which indicates that the TAs and GW do not agree on how to grade the individual grading items. Since GW made no known errors and instance where the TAs and GW disagreed were spot checked, the grading by GW will be treated as the ground truth. A potential problem with this method is if both the TAs and GW incorrectly graded at the same time (i.e., they both had simultaneous false negatives) then this type of error would not be detected. However, since GW is only a computer algorithm and the course instructor found no mistakes with its grading, simultaneous false negatives is unlikely. As a result, the 41 grading items with p-value lower than 0.05 is interpreted as the TAs making a statistically significant amount of errors.

The results show that some grading items are significantly more manageable for a human TA to consistenly grade accurately than others. In the following paragraphs, the results are discussed based on the category of the grading items (volume, material, fully defined sketches, and shape).

For all three assignments, the TAs were able to correctly grade the volume of the parts 97% of the time. On assignment three, the volume grading item on the TA rubric was incorrectly labeled as "mass," however, the units were for volume (see Appendix A). Figure 8 shows that three of the TAs were able to catch this mistake, but one grader did not find the error and marked many correct submissions as incorrect.

The TA graders were also able to correctly grade the material 99% of the time. However, again the rubric on assignment two that the TAs fill out had the wrong material for one of the parts. Figure 8 shows that grader one did not catch the mistake while the other graders correctly graded what the rubric should have listed. The high agreement between the TAs and GW for the material also increases our confidence that the software is correctly identifying which student file matches with which rubric item and that the software is correctly applying its internal rubric to the student's files. For the TAs, grading the material assigned to the part file is straighforward as it is immediately visible on the feature tree when you open a part in SOLIDWORKS.

The TAs were able to correctly check for students fully defined sketches $\approx$ 92% of the time when considering assignments 3 and 4. The rubric for assignment 2 combined checking for fully defined sketches and checking the shape of the part, an essential part of 3D modeling. Both the TAs and GW show that on assignment four, only about 65% of students fully defined their sketches. This is a likely cause of volume and shape errors since the student's sketches are still vague about the actual shape of the part. One advantage of GW is that it lists the sketches which are under-defined when filling out the rubric. Likely, listing the specific problematic sketches would help the students realize the importance of fully defining their sketches.

Visually checking for fully defines sketches in SOLIDWORKS requires checking the feature tree icon for each sketch. Each sketch that is not fully defined will have a dash (-) in front of it, but some of the TAs did not know this shortcut and were opening each sketch to check if it was fully defined. Training TAs to use this shortcut could improve their performance on this rubric item.

The shape of the parts presents the most significant challenge for the TA graders. The assignments become progressively more complex. Each part is based on hundreds of parameters such as sketch relations, dimensions, and feature parameters. For the TA graders, checking each parameter is not realistic since they must grade so many files in a limited time frame. As a result, the rubrics have been designed to check the "general appearance and shape." Significant deviations from the correct answer model are typically found by the TA graders, however, small mistakes were often overlooked.

Maximizing consistency and accuracy is desired to ensure a quality learning experience and fairness across students. Simply having a “correct” volume or mass does not definitively prove a part has the correct shape. For example, Figure 12 shows that for the three advanced modeling parts in assignment four, the volume of the object was often within 1% of the correct answer yet the shape was still incorrect; volume is only a fair proxy for accuracy. The CSWA exam administered by SOLIDWORKS uses the mass (which is a function of assigning the proper material and having the exact volume) to grade user’s part files. The results of this study suggest grading based on just the mass is reliable, but not always valid and should be done with caution or in combination with other measures to ensure the most reliable and valid measurement. However, the inconsistency between graders indicates potential problems with both reliability and validity when grading is completed without software assistance even when a rubric is used. To evaluate the difference between graders, a two-way ANOVA test was performed. The input is the agreement (1) or disagreement (0) with GW. The grouping is by type of grading (material, volume, shape, sketches), and the treatment is the different graders. The null hypothesis is that the TAs perform similarly on similar types of grading tasks. In other words, the null hypothesis expects that they make mistakes at the same rate, and as a result the average number of errors (disagreement with GW) is the same for each grader.
Table *3* shows the F-value and p-values from the ANOVA test. All the p-values are well below 0.05 which means that we can reject the null hypothesis. Significant variation does exist between graders and for different types of grading items. To better understand why the null hypothesis is rejected, confidence interval plots for $\alpha = 0.05$ using Tukey-Kramer criterion are shown in figures 14, 15, and 16. The plot axes are normalized where one represents 100% agreement and zero represents 0% agreement. The circles in the plots represent the mean number of agreements with GW.

Table 3. Analysis of variation between TA graders and grading type.

Assignment	Variation in Grader (F-value)	Variation in Grading type (F-value)	Variation in Grader (p-value)	Variation in Grading type (p-value)
A2	54.8	169.5	1.0e-34	3.3e-71
A3	20.1	437.2	5.4e-13	1.3e-264
A4	11.3	1678.3	1.8e-07	0

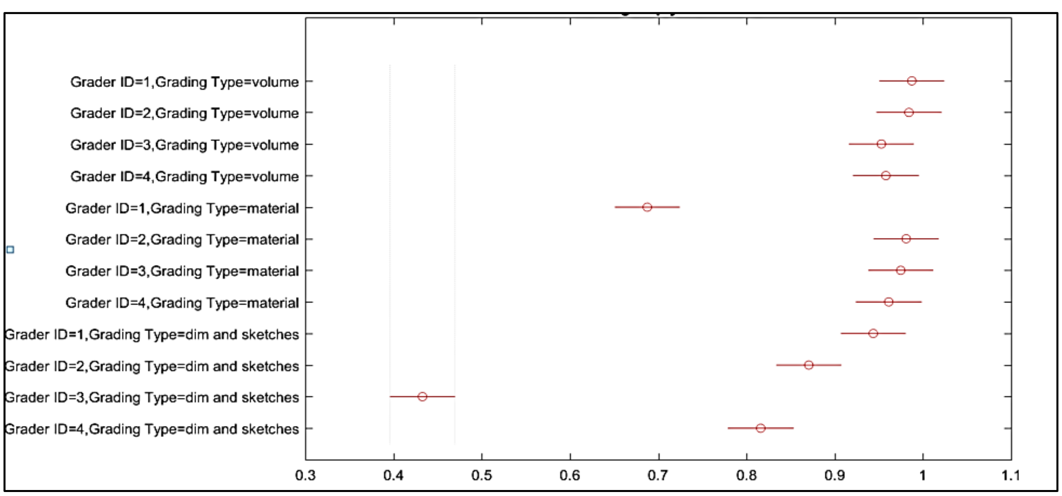


Figure 14. A2 analysis of variation for combinations of groups and treatments with p<= 0.05. The circles in the plots represent the mean number of agreements with GW.

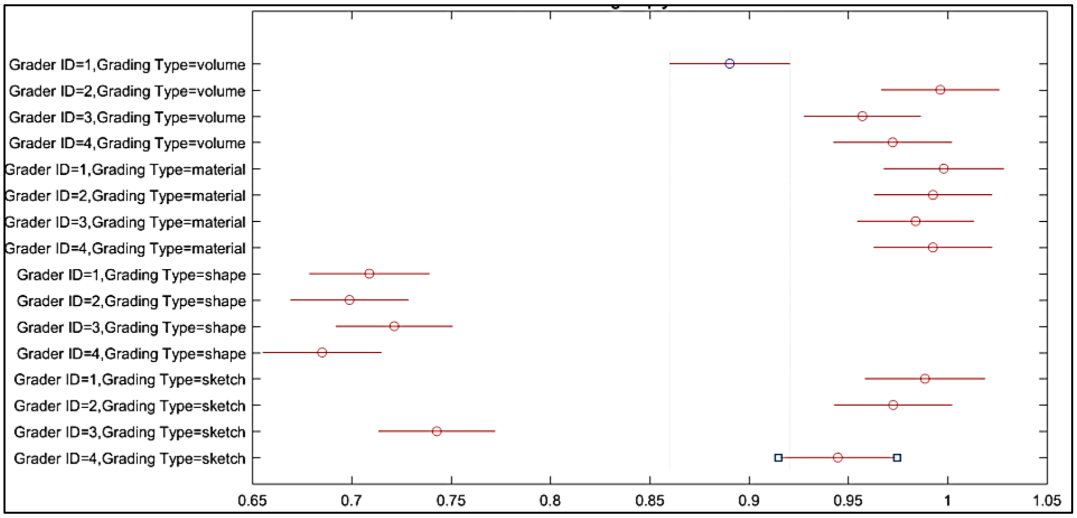


Figure 15. A3 analysis of variation for combinations of groups and treatments with p <= 0.05. The circles in the plots represent the mean number of agreements with GW.

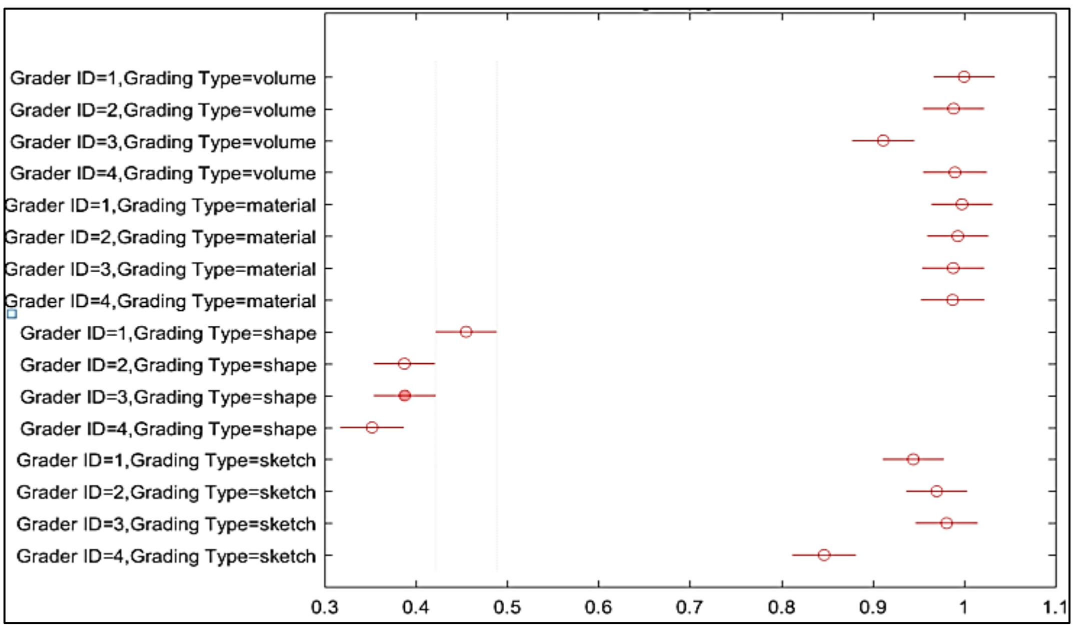


Figure 16. A4 analysis of variation for combinations of groups and treatments with p <= 0.05. The circles in the plots represent the mean number of agreements with GW.

The confidence interval plots show that the mean of most of the group and treatment combinations fall between 0.9 and 1.0 and that the confidence intervals overlap. However, grading the shape of a 3D model results in a much higher mean number of errors for all graders although this mean still has statistically significant difference for different graders on A4. When designing the TA rubrics, one of the objectives was ease of use to foster consistent grading, yet results indicate that the graders are not consistent in the number or types of errors they make. In figures 7, 10, and 13, the histogram of final grades for each assignment shows that both GW and the TAs results yield similar grade distribution. The GW distribution is more nuanced than the TA grading because GW assigns partial credit on the shape grading item using a computational shape similarity score. A student could earn a 95% on their GW shape, but the shape score on the TA rubric is pass or fail with no partial credit. This result strengthens the finding of Renu et al. that using shape similarity algorithms can result in a better distribution of grades than simple pass or fail evaluations [9].

Discussion

The results indicate that GW is an effective grading support software for SOLIDWORKS part files when grading simple characteristics of student's files such as the material, volume, shape, and checking for full defined sketches. Therefore, a strength of computational grading tools is their ability to apply simple grading criteria to the student files consistently. In contrast, the TA made many mistakes and were not consistent between graders, though overall, grade distributions were similar between the two methods.

Correctly grading student's assignments is essential as students may not realize that they made a mistake if they do not lose a few points on the assignments for those mistakes. The entire purpose for grading assignments is so students can learn from their previous mistakes and overcome skills deficiencies. Correctly grading student's assignment is also essential for assessment. On assignment 4, the TA graders were not able to identify very small mistakes in the student's files. If the graders identified the top few common mistakes, then the course instructors could have used this information to reteach the concepts during lecture.

While the software was able to grade files more accurately and quicker than the TA graders, the software has limitations. In this study neither the TAs nor GW attempted to evaluate the student's models on a higher level of abstraction such as assessing design intent or offering specific feedback on what the student did incorrectly. Developing computational tools to grade design intent or provide specific feedback would be difficult since there are many different ways to create equivalent 3D CAD models. Additionally, Graderworks' shape similarity score is a black box with no description of how it works and the shape score is slightly stochastic. Both of these features may cause some instructors to distrust the Graderworks output.

Small mistakes based on incorrectly specifying a single parameter in a part file can significantly change the shape. GW is not able to identify that only a single parameter is incorrect and therefore may take off many points on the shape score. The TA graders could be trained to identify what a student did wrong and offer constructive feedback. GW is better able to detect that a part file has a mistake or does not have a mistake.

GW is helpful for grading part files and rigid assemblies, but an engineering graphics course covers many more topics that are important for students to know. Other important topics include engineering drawings, moving assemblies, FEA, photoview rendering, and animation. All of these still require some human grading.

Conclusions

This work compared software grading to graduate student TA grading for three separate assignments in an engineering graphics class with about 450 students. The GW software proved to be a more reliable and valid method of grading part files than graduate student teaching assistants on many grading aspects: material, volume, checking for fully defined sketches, and checking the shape of the student's part files.

The most significant difference in the TA graders and the software is when grading the shape of complex parts. These complex parts are based on hundreds of parameters such as the sketch relations, dimensions, and feature parameters and dimensions. Since the parts are complex, it is not realistic for a human TA to check every single feature, but instead, the TA's check the general shape compared to the correct answer's shape. During this comparison, the TA's can identify significant mistakes made by the students, but smaller mistakes often go unnoticed. For the three most complex parts in this study, the TA graders incorrectly graded the shape of these parts $\approx 70\%$ of the time.

One of the authors spot checked the part files that GW identified as having almost the correct shapes. Initially finding the student's mistake was difficult and took several minutes; however, after common mistakes that slightly changed the shape were identified finding student's mistakes was much easier. This results suggest that assessing if two models have the exact same shape is a very difficult task for human graders whether it is a TA or an instructor.

The results from this study also showed that volume or mass of part is often a flawed proxy for evaluating accuracy of part shape. This result is a bit concerning as it is the method used by the primary external standard for judging mastery of SOLIDWORKS, the CSWA exam.

One software limitations is that is cannot determine exact cause for the error on their part file. Because the software is better grading basic characteristics of the student's files than the TA graders, the TAs could be retrained to provide better feedback in conjunction with the feedback from GW. GW could indicate which files have errors. The TAs could look at these files and offer specific feedback to the students on what they did wrong or a better way to model the part or assembly. In other words, GW could free from the basic assessments they perform now so that they could offer a higher level of feedback to the student which would likely increase achievement of student learning outcomes.

Future work

Immediate next steps will attempt to fuse the two methods of grading into the best combination. If GW is used to form a base evaluation of student models and identify those containing major errors, then the TAs could focus on providing targeted feedback regarding what was done wrong or provide alternative design strategies. Although it seems likely that student outcomes would improve, research is needed to identify the optimal cutoff threshold for prioritizing a submission for TA review.

Future work will evaluate the use of GW software as a method of identifying instances of academic integrity violations such as submitting a file that was created in a prior academic semester. Students are expected to complete their own work and are on their honor to do so. Human graders typically only identify a few (if any) potentially concerning submissions each semester. It is expected that GW could increase the number of instances identified and provide evidence to support the suspicion. Else, informing students of the use of such software tools may ultimately reduce the number of instances of violations.

Potential software development work will expand GW to applicable for other 3D CAD software besides SOLIDWORKS. Development of tools similar to GW for other 3D CAD software would be helpful to many academic institutions and may have industry applications as well.

Disclosure

Anthony Garland has an ownership stake in Garland Industries LLC which develops Graderworks.

References

[1] "SOLIDWORKS," Dassault Systems, [Online]. Available: https://www.3ds.com/products-services/solidworks/ . [Accessed 5 May 2018].

[2] A. Reyes, Beginner's Guide to SOLIDWORKS 2018 - Level I, Mission, KS, USA: SDC Publications, 2017, p. 752.

[3] Ault, H. K., & Fraser, A. (June, 2013). A comparison of manual vs. online grading for solid models. Proceedings of the 2013 ASEE Annual Conference, Atlanta, Georgia, June 23-26, 2013.

[4] A. Rynne, W. F. Gaughran and B. McNamara, "Parametric modelling training strategies to capture design intent," in *ICPR-17th International Conference on Production Research conference proceedings*, 2003.

[5] P. Company, M. Contero, J. Otey and R. Plumed, "Computer-Aided Design Approach for developing coordinated rubrics to convey quality criteria in MCAD training," *Computer-Aided Design*, vol. 63, pp. 101-117, 2015.

[6] C. González-Lluch, P. Company, M. Contero, J. Camba and J. Colom, "A case study on the use of model quality testing tools for the assessment of MCAD models and drawings," *International Journal of Engineering Education*, vol. 33, no. 5, pp. 1643-1653, 2017.

[7] M. Hekman,K.A.; Gordon, "Automated Grading of First Year Student CAD Work,," *120th ASEE Annual Conference & Exposition*, pp. 23-26, Paper ID #6379., 2013.

[8] S. Ingale, S. Anirudh and D. Bairaktarova, "CAD Platform Independent Software For Automatic Grading Of Technical Drawings (DETC2017-67612)," in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Cleveland, Ohio, USA, 2017.

[9] R. Renu, A. Garland, S. Grigg, J. Minor and N. Yasmin, "Improving Engineering Graphics Grading Using a Shape Similarity Algorithm– An Initial Investigation," in *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering*, 2016.

[10] G. Paravati, S. Member, F. Lamberti and S. Member, "Point Cloud-Based Automatic Assessment of 3D Computer Animation Courseworks," vol. 10, no. 4, pp. 532-543, 2017.

[11] Kirstukas, S. J. (2016, June), Development and Evaluation of a Computer Program to Assess Student CAD Models Paper presented at 2016 ASEE Annual Conference & Exposition, New Orleans, Louisiana. 10.18260/p.26781

[12] "SOLIDWORKS Certification Center," [Online]. Available: https://solidworks.virtualtester.com.

[13] A. Cardone, S. K. Gupta and M. Karnik, "A Survey of Shape Similarity Assessment Algorithms for Product Design and Manufacturing Applications," *Journal of Computing and Information Science in Engineering*, vol. 3, no. 2, p. 109, 2003.

Appendix A: TA Rubrics

Assignment	Name	Points Possible	Points Lost	Late penalty	Grade
A2		100	0		100/100
	Description	Possible	Lost	Comments	
Formatting	File Submission A2 (10)	10			
	Files Named Correctly	3			
	Folder named correctly (username-A2)	3			
	Submitted .zip folder of parts	4			
Exercise 4	Part Submission Exercise 4 (30)	30			
	Linear pattern square indents (or equivalent)	5			
	Slot extrusion (or equivalent)	5			
	Correctly dimensioned, fully defined	5			
	Units: mmgs	5			
	Assigned material (ABS)	5			
	Volume: *232908.7385 cubic millimeters* (235237.8259 - 230579.6511)	5			
Circular Plate	Part Submission Circular Plate (30)	30			
	Hexagon hole	4			
	Circular pattern (or equivalent)	3			
	Filleted corners on holes (different top and bottom)	3			
	Correctly dimensioned, fully defined	5			
	Units: MMGS	5			
	Assigned material (Cast Alloy Steel)	5			
	Volume: 89038.1491 cubic millimeters (89928.5306 - 88147.7676)	5			
Slotted Plate	Part Submission Slotted Plate (30)	30			
	Two slots	5			
	Tangent relation between arc and line	5			
	Correctly dimensioned, fully defined	5			
	Units: IPS	5			
	Assigned material (Plain Carbon Steel)	5			
	Volume: *15.9429 cubic inches* (16.1023 - 15.7834)	5			

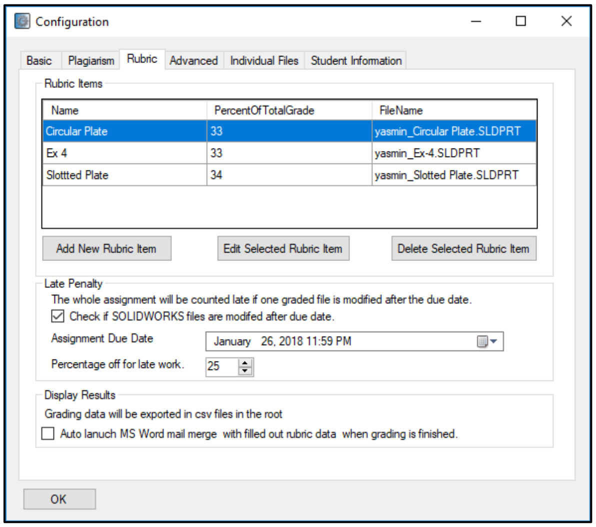
Assignment	Name	Points Possible	Points Lost	Late penalty	Grade
A3		100	0		100/100
	Description	Possible	Lost	Comments	
Formatting	Formatting (10)	10			
	Files Named Correctly	3			
	Folder named correctly (username-A3)	3			
	Submitted .zip folder of parts	4			
Housing	Housing (22)	22			
	Volume: 19.1782 cu in (19.370 - 18.986)	3			
	Units: IPS	3			
	Material: Cast Alloy Steel	2			
	General Appearance and Shape	2			
	Fully defined sketches	2			
	Base Extrusion	2			
	Inside cut	2			
	Fillets	2			
	Circle boss mirrored	2			
	Side boss mirrored	2			
Cylinder Gasket	Cylinder Gasket (15)	15			
	Mass: 0.207 cu in (0.209 - 0.205)	3			
	Units: IPS	2			
	Material: Viton	2			
	General Appearance and Shape	2			
	Fully defined sketches	2			
	Fillets or rounded corners	2			
	Holes	2			
Side Cover	Side Cover (17)	17			
	Volume: 1.272 cu in (1.285 - 1.259)	3			
	Units: IPS	2			
	Material: AISI 1020	2			
	General Appearance and Shape	2			
	Fully defined sketches	2			
	Center hub	2			
	Six Holes	2			
	Fillet around hub	2			
Exercise 5	Exercise 5 (19)	19			
	Volume: 10.9094 cu in (11.018 - 10.800)	3			
	Units: IPS	2			
	Material: 6061 Alloy (Al)	2			
	General Appearance and Shape	2			
	Fully defined sketches	2			
	Base Revolve or equivalent	2			
	Outer circular pattern of holes	2			
	Inner circular pattern of holes	2			
	Center slot	2			
Top Cover	Top Cover (17)	17			
	Volume: 1.8225 cu in (1.8407 - 1.8043)	3			
	Units: IPS	2			
	Material: Cast Alloy Steel	2			
	General Appearance and Shape	2			
	Fully defined sketches	2			
	Fillets	2			
	Shell	2			
	Hole wizard holes	2			

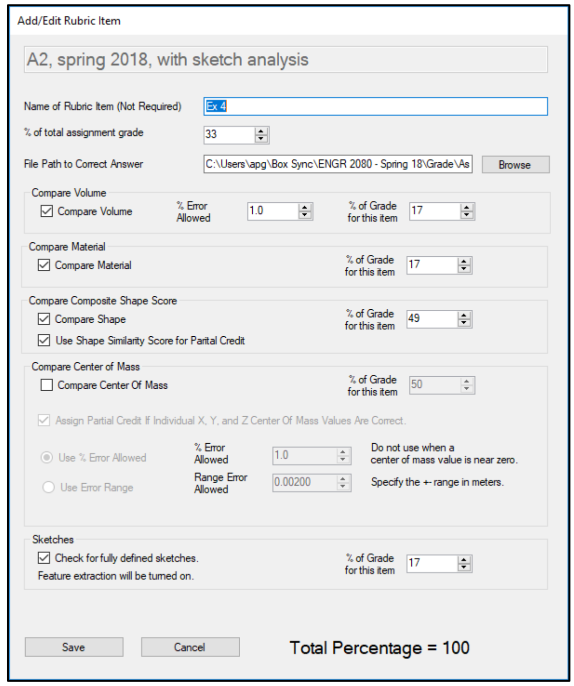
Assignment	Name	Points Possible	Points Lost	Late penalty	Grade
A4		100	0		100/100

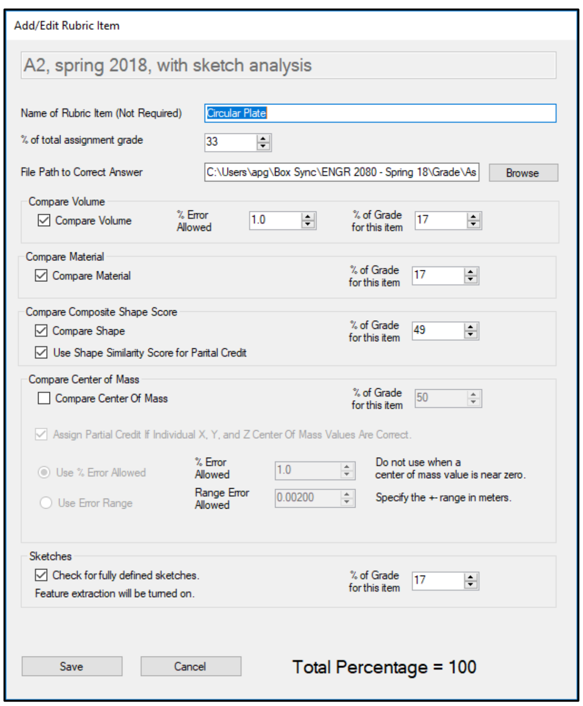
		Possible	Lost	Comments
Formatting	Formatting (10)	10		
	Files Named Correctly	3		
	Folder named correctly (username-A4)	3		
	Submitted .zip folder of parts	4		
Worm Gear Shaft	Worm Gear Shaft (21)	21		
	Volume: 1.8594 cu. in (1.878 - 1.8408)	3		
	Units: IPS	3		
	Material: Chrome Stainless Steel	3		
	General Appearance and Shape	3		
	Fully defined sketches	3		
	Base Extrusion	2		
	Offset grooves / Hex endcap	2		
	Keyway cut	2		
Offset Shaft	Offset Shaft (23)	23		
	Volume: 1.9036 cu. in (1.9226 - 1.88466)	3		
	Units: IPS	3		
	Material: Chrome Stainless Steel	3		
	General Appearance and Shape	3		
	Fully defined sketches	3		
	Base Extrusions	2		
	Offset grooves / Hex endcap	2		
	Thread sweep with Tapered Helix	2		
	Fillet around thread	2		
Worm Gear	Worm Gear (23)	23		
	Volume: 1.1493 cu. in (1.1608 - 1.1378)	3		
	Units: IPS	3		
	Material: AISI 1020	3		
	General Appearance and Shape	3		
	Fully defined sketches	3		
	Extrusion and Revolved cut or equivalent	2		
	Keyway cut/Fillets	2		
	Cut Sweep teeth	2		
	Circular Pattern teeth (or equivalent)	2		
Housing	Housing (23)	23		
	Volume:16.6169 cu. in (16.7831 - 16.4507)	3		
	Units: IPS	3		
	Material: Cast Alloy Steel	3		
	General Appearance and Shape	3		
	Fully defined sketches	3		
	Hole wizard holes	2		
	Dimension Modifications	2		
	Circular Pattern/Mirror (Screw holes)	2		
	Slots / Fillets	2		

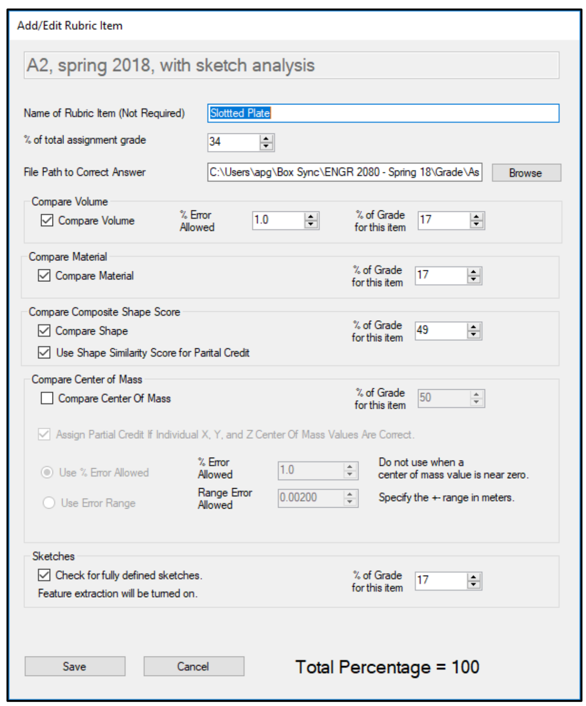
Appendix B : GW Rubrics

Assignment 2 Rubrics

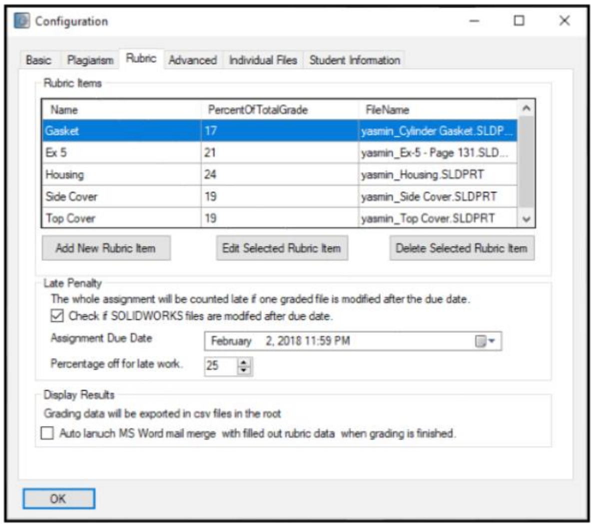


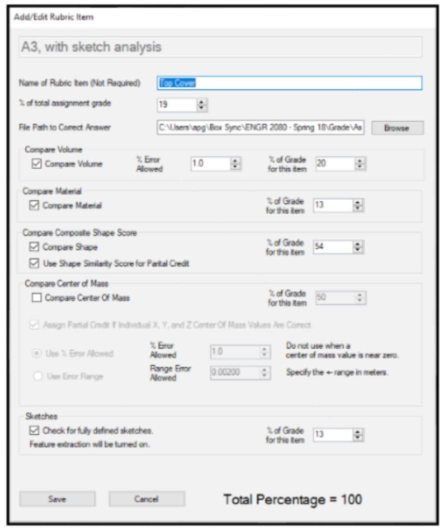


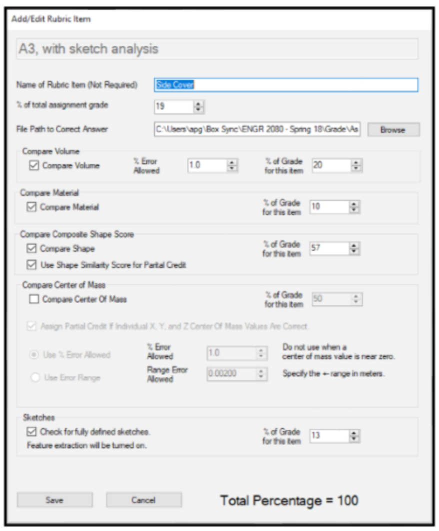


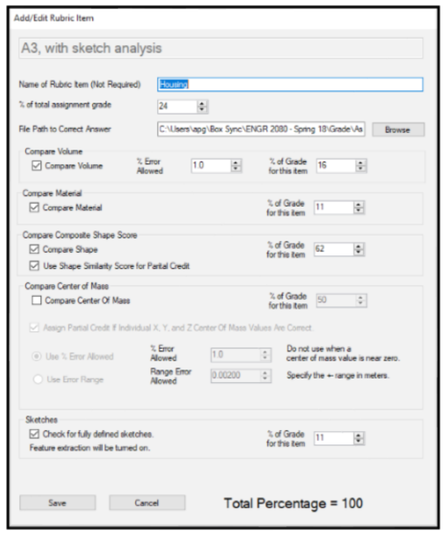


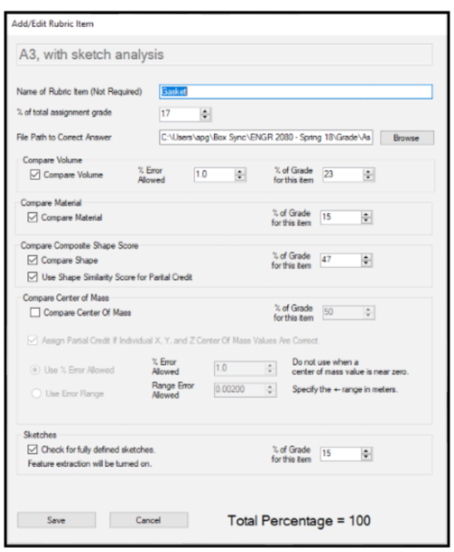
Assignment 3. GW rubrics

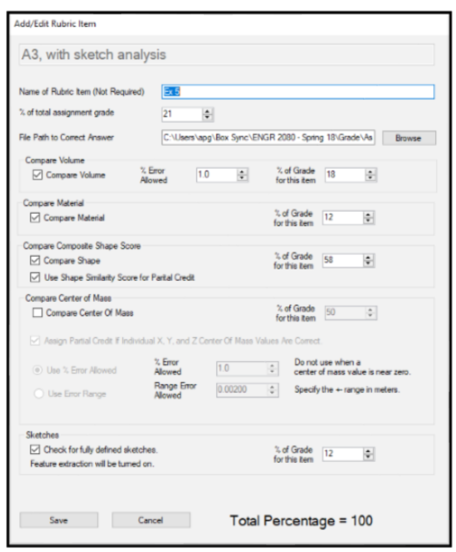




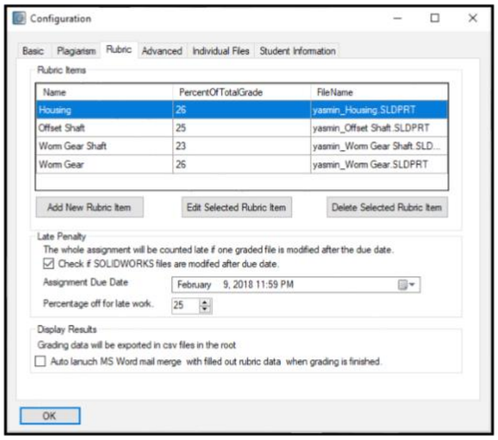


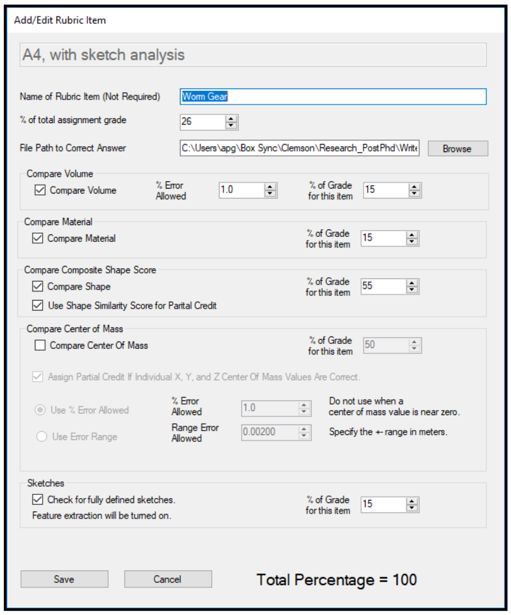


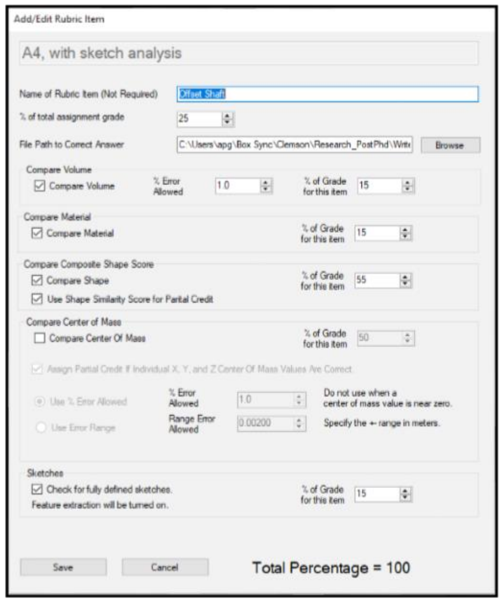


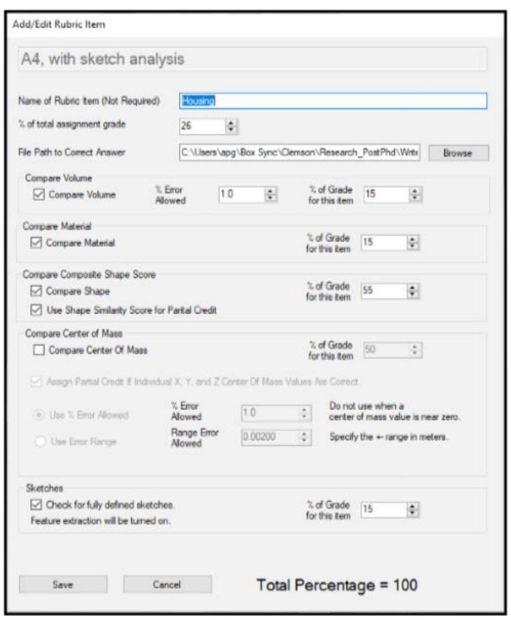


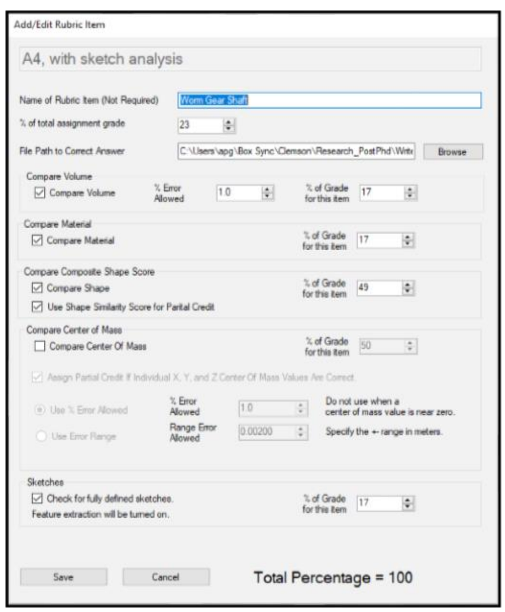
Assignment 4 GW rubrics











Appendix C: Itemized comparison of TA and GW

Note: % agreement and % disagreement means agreement or disagreement of the TA with GW. ** (TA rubric had wrong material)

Assignment 2

	All Graders			Grader 1			Grader 2			Grader 3			Grader 4		
Rubric item Name	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree
Volume Of Circular Plate	432	96.53	3.47	110	96.53	3.47	107	96.53	3.47	109	96.53	3.47	106	96.53	3.47
Material Circular Plate **	432	73.61	26.39	110	73.61	26.39	107	73.61	26.39	109	73.61	26.39	106	73.61	26.39
Compare dim and sketches circular plate	432	74.54	25.46	110	74.54	25.46	107	74.54	25.46	109	74.54	25.46	106	74.54	25.46
Volume Of Ex4	419	97.61	2.39	106	97.61	2.39	105	97.61	2.39	105	97.61	2.39	103	97.61	2.39
Material Of Ex4	419	98.57	1.43	106	98.57	1.43	105	98.57	1.43	105	98.57	1.43	103	98.57	1.43
Compare dim and sketches Of Ex4	419	73.75	26.25	106	73.75	26.25	105	73.75	26.25	105	73.75	26.25	103	73.75	26.25
Volume Slotted Plate	415	97.11	2.89	104	97.11	2.89	105	97.11	2.89	105	97.11	2.89	101	97.11	2.89
Material Slotted Plate	415	98.55	1.45	104	98.55	1.45	105	98.55	1.45	105	98.55	1.45	101	98.55	1.45
Compare dim and sketches of Slotted Plate	415	81.45	18.55	104	81.45	18.55	105	81.45	18.55	105	81.45	18.55	101	81.45	18.55

Assignment 3

**(TA rubric listed as mass)

	All Graders			Grader 1			Grader 2			Grader 3			Grader 4		
Rubric item Name	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree
Volume gasket **	437	82.61	17.39	107	82.61	17.39	110	82.61	17.39	112	82.61	17.39	108	82.61	17.39
Material gasket	437	99.08	0.92	107	99.08	0.92	110	99.08	0.92	112	99.08	0.92	108	99.08	0.92
Shape of gasket	437	72.31	27.69	107	72.31	27.69	110	72.31	27.69	112	72.31	27.69	108	72.31	27.69
Sketches of gasket	437	94.28	5.72	107	94.28	5.72	110	94.28	5.72	112	94.28	5.72	108	94.28	5.72
Volume Ex 5	432	98.61	1.39	104	98.61	1.39	109	98.61	1.39	112	98.61	1.39	107	98.61	1.39
Material Ex 5	432	98.38	1.62	104	98.38	1.62	109	98.38	1.62	112	98.38	1.62	107	98.38	1.62
Shape Ex 5	432	67.59	32.41	104	67.59	32.41	109	67.59	32.41	112	67.59	32.41	107	67.59	32.41
Sketch Ex 5	432	87.96	12.04	104	87.96	12.04	109	87.96	12.04	112	87.96	12.04	107	87.96	12.04
Volume Housing	439	99.09	0.91	108	99.09	0.91	109	99.09	0.91	112	99.09	0.91	110	99.09	0.91
Material Housing	439	99.77	0.23	108	99.77	0.23	109	99.77	0.23	112	99.77	0.23	110	99.77	0.23
Shape Housing	439	52.16	47.84	108	52.16	47.84	109	52.16	47.84	112	52.16	47.84	110	52.16	47.84
Sketches Housing	439	94.08	5.92	108	94.08	5.92	109	94.08	5.92	112	94.08	5.92	110	94.08	5.92
Volume Side Cover	437	99.08	0.92	105	99.08	0.92	110	99.08	0.92	112	99.08	0.92	110	99.08	0.92
Material Side Cover	437	98.86	1.14	105	98.86	1.14	110	98.86	1.14	112	98.86	1.14	110	98.86	1.14
Shape Side Cover	437	90.39	9.61	105	90.39	9.61	110	90.39	9.61	112	90.39	9.61	110	90.39	9.61
Sketches Side Cover	437	89.47	10.53	105	89.47	10.53	110	89.47	10.53	112	89.47	10.53	110	89.47	10.53
Volume Top Cover	435	97.93	2.07	105	97.93	2.07	110	97.93	2.07	112	97.93	2.07	108	97.93	2.07
Material Top Cover	435	99.77	0.23	105	99.77	0.23	110	99.77	0.23	112	99.77	0.23	108	99.77	0.23
Shape Top Cover	435	69.43	30.57	105	69.43	30.57	110	69.43	30.57	112	69.43	30.57	108	69.43	30.57
Sketches Top Cover	435	89.43	10.57	105	89.43	10.57	110	89.43	10.57	112	89.43	10.57	108	89.43	10.57

Assignment 4

	All Graders			Grader 1			Grader 2			Grader 3			Grader 4		
Rubric item Name	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree	N	% agree	% disagree
Volume Housing (A4)	424	97.88	2.12	108	97.88	2.12	109	97.88	2.12	105	97.88	2.12	102	97.88	2.12
Material Housing (A4)	424	99.53	0.47	108	99.53	0.47	109	99.53	0.47	105	99.53	0.47	102	99.53	0.47
Shape Housing (A4)	424	32.55	67.45	108	32.55	67.45	109	32.55	67.45	105	32.55	67.45	102	32.55	67.45
Sketches Housing (A4)	424	92.69	7.31	108	92.69	7.31	109	92.69	7.31	105	92.69	7.31	102	92.69	7.31
Volume Offset Shaft (A4)	422	97.39	2.61	108	97.39	2.61	109	97.39	2.61	105	97.39	2.61	100	97.39	2.61
Material Offset Shaft (A4)	422	99.53	0.47	108	99.53	0.47	109	99.53	0.47	105	99.53	0.47	100	99.53	0.47
Shape Offset Shaft (A4)	422	33.41	66.59	108	33.41	66.59	109	33.41	66.59	105	33.41	66.59	100	33.41	66.59
Sketches Offset Shaft (A4)	422	92.18	7.82	108	92.18	7.82	109	92.18	7.82	105	92.18	7.82	100	92.18	7.82
Volume Worm Gear Shaft (A4)	415	97.59	2.41	107	97.59	2.41	108	97.59	2.41	103	97.59	2.41	97	97.59	2.41
Material Worm Gear Shaft (A4)	415	98.55	1.45	107	98.55	1.45	108	98.55	1.45	103	98.55	1.45	97	98.55	1.45
Shape Worm Gear Shaft (A4)	415	60.96	39.04	107	60.96	39.04	108	60.96	39.04	103	60.96	39.04	97	60.96	39.04
Sketches Worm Gear Shaft (A4)	415	98.55	1.45	107	98.55	1.45	108	98.55	1.45	103	98.55	1.45	97	98.55	1.45
Volume Worm Gear (A4)	420	96.19	3.81	108	96.19	3.81	108	96.19	3.81	105	96.19	3.81	99	96.19	3.81
Material Worm Gear (A4)	420	99.05	0.95	108	99.05	0.95	108	99.05	0.95	105	99.05	0.95	99	99.05	0.95
Shape Worm Gear (A4)	420	31.90	68.10	108	31.90	68.10	108	31.90	68.10	105	31.90	68.10	99	31.90	68.10
Sketches Worm Gear (A4)	420	91.43	8.57	108	91.43	8.57	108	91.43	8.57	105	91.43	8.57	99	91.43	8.57