Testing Inter-Rater Reliability in Rubrics for Large Scale Undergraduate Independent Projects

Alan Chong and Lisa Romkey
University of Toronto
alan.chong@utoronto.ca; lisa.romkey@utoronto.ca

***Abstract** –This work outlines the process of testing inter-rater reliability in rubrics for large scale undergraduate independent projects; more specifically, the thesis program within the Division of Engineering Science at the University of Toronto, in which 200 students work with over 100 supervisors on an independent research project. Over the last few years, rubrics have been developed to both guide the students in the creation of their thesis deliverables, and to improve the consistency of supervisor assessment. To examine inter-rater reliability, 12 final thesis reports were assessed using the course rubric by the two generalist experts, who have worked extensively with the thesis course and designed the rubrics, alongside the project supervisor. We found substantial agreement between the two generalist experts, but only fair agreement between the generalist experts and the supervisors, suggesting that while the rubric does help towards developing a common set of expectations, there may be other aspects of the supervisor's assessment practice that need to be considered.*

***Keywords:* Rubrics, Inter-rater reliability, Capstone, Assessment**

1. INTRODUCTION

Designing and validating rubrics for capstone experiences, such as fourth year design projects and research-oriented theses, is a long and challenging process, involving attempts to balance the needs of a department, students, as well as supervisors, each of whom have different priorities. In a large undergraduate program, such as Engineering Science at the University of Toronto, a required thesis course essentially amounts to 200 independent study courses, necessitating strategies – such as universal rubrics – to ensure some degree of uniformity of experience and expectations between a large number of supervisors. Many of these supervisors, especially those outside the Faculty of Applied Science and Engineering, are not used to working with Engineering Science students and have asked for guidance in assessment. Students, on the other hand, have used the rubrics as guidelines for both their deliverables and managing the research experience overall.

Previous papers have discussed the challenges involved in rubric development in this context, particularly those presented by the variety of projects and expectations of instructors involved in assessment of student deliverables, as well as the use of focus groups to gain a deeper understanding of student perspectives on course deliverables and rubric criteria [2, 3, 13]. This research has led to revisions in the rubrics deployed in the course which, we hypothesize, best balance the needs of all stakeholders in the process.

One key test towards validating this hypothesis involves measuring the scoring consistency or inter-rater reliability facilitated by the rubric. A traditional test would employ multiple expert assessors on a large sample of documents to measure inter-rater reliability. However, for this capstone experience, such a methodology would be difficult for two reasons. First of all, the thesis documents are highly specialized and discipline specific, meaning few expert raters could be found. Second, the documents range from 50-100 pages, meaning many assessments would necessarily be highly resource intensive.

For these reasons, in this study, two generalist expert raters – faculty members involved in supporting the thesis experience for all students and responsible for developing the rubric – assessed 12 theses and compared their results with each other, as well as the official scoring rubric of the actual supervisor. In addition to providing an assessment of inter-rater reliability, this process elicited some new insights about the design and use of the rubric, which are reported on in this paper.

2. BACKGROUND

2.1. Previous Work

While the required thesis course is a long-standing curriculum component in the Engineering Science program at the University of Toronto, it is only in the last six years that the course has employed instruction, assignment guides and rubrics to support the written and oral deliverables associated with the course, and the research experience as a whole. This change was the result of both students and supervisors expressing a need for better guidelines to help shape the thesis experience, and an interest in facilitating common assessment practices

and inter-rater reliability across a large number of supervisors. An added requirement emerged in 2012, with the introduction of the graduate attributes process, and the thesis rubrics were then adapted for use in the assessment of the communication and investigation attributes [2].

As described in [2] and [3], our thinking around inter-rater reliability and the challenge of testing it was influenced by two mitigating factors: first, the students are assessed by full faculty members who are experts in their field. Second, the way in which the rubrics have been designed - including a "rough guide" with specific written descriptions of each requirement at each level of performance - was an attempt to create clarity and a common vocabulary across assessors, which has also been explored in other research as a way of improving inter-rater reliability [1, 11]. However, despite the challenging circumstances for measuring inter-rater reliability, it is an important step in continuing to evolve and improve the rubrics associated with the course.

In previous work, the student interpretation of the rubric requirements and the applicability of the rubrics to individual student projects, through the use of student focus groups, has been described [13]. This work elicited a better understanding of how the students use the rubric, vocabulary in the rubric that was unclear, and the students' impression of how their supervisor was (or perhaps, was not) using the grading rubric.

The criteria used in the rubric are described below, in Table 1, each of which is scored with four levels of achievement; poor, adequate, good, and exceeds expectations. The first four categories correspond to different sections of the document, or rhetorical tasks the thesis must perform, with each criteria setting out goals for that section. These categories represent an attempt to balance brevity, usability, and informativeness for supervisors and students. As mentioned in [2], we have also produced a Rough Guide to the Thesis Requirements, shown in Appendix A, in response to supervisors and students seeking a more detailed explanation of what each of the four levels of achievement mean with respect to each criterion.

Table 1. Engineering Science Thesis Final Document Rubric Requirements

Introduction/ Background	Establishes context necessary to facilitate thorough understanding of thesis work in a concise manner
	Establishes a clear research gap/design problem, makes a convincing case for the significance of proposed research work
	Identifies goal for thesis work that explicitly addresses this gap/problem; provides clear purpose statement

Literature Review/ Background	Explains theoretical concepts important to understanding of thesis work
	Identifies, summarizes, and synthesizes relevant research in constructing an understanding of current state of field
	Enables deeper understanding of research question/design problem through analysis of research in the field, indicating a path for moving research forward
Methods and Findings	Describes methods or design in sufficient detail to enable understanding of work done
	Provides justification for methods chosen or design decisions made
	Results displayed clearly in organized manner, using appropriate figures or graphics; key results highlighted
Discussion and Conclusions	Engages with and explains results intelligently
	Identifies key claims to be drawn from results of research or design evaluation, qualifies them appropriately
	Outlines significance of research done, identifies potential future work that arises from thesis work
Overall/ Document Design	Abstract concisely summarizes purpose, methods, key results of research, and presents conclusions clearly
	Document length, formatting, structure meets stated requirements, and specific demands of thesis topic
	Organized well, with content in discrete and appropriate positions in paper, structure clearly laid out, transitions that create flow in document
	Demonstrates grammatically correct, coherent prose that concisely and clearly communicates complex topics in well designed paragraphs and sentences
	Uses and integrates well-designed visuals effectively to communicate key concepts / results
	Provides clear attribution of ideas throughout paper using a known referencing standard; uses references effectively to help establish context, back claims, or justify decisions

The rubric also includes another section, with six criteria, describing "Project Experience." However, that section describes components such as whether the student "demonstrated initiative, an ability to work independently,

time management skills and ownership of work throughout thesis project", alongside other aspects of the student's performance over the course of the year, measured through interactions with the supervisor. We chose, therefore, to remove those components from our analysis.

2.2 Literature on Inter-rater Reliability

Inter-rater reliability represents the extent to which different reviewers assign the same score to a particular variable - in this case, a requirement on a rubric. Although in ideal circumstances, a large number of raters may be used to measure inter-rater reliability, it is often the case, particularly in unique capstone experiences, that it simply isn't possible. As noted by Jonsson and Svingby in a review of 75 papers about scoring rubrics, "two raters are, under restrained conditions, enough to produce acceptable levels of inter-rater agreement" [8]. This review paper also examined the specific calculations used to measure inter-rater reliability across these 75 papers, and found that a large number of studies use percentage agreement (exact or adjacent), and/or a measure of correlation such as Cohen's kappa, Pearson's r or Cronbach's alpha. What is deemed as an acceptable level of inter-rater reliability, across this literature review, depended on the type of assessment; for example, consider formative vs. high-stakes summative, and the importance of a particular level of inter-rater reliability in each case. In examining a few specific examples of similar studies with a small number (2-7) of raters focusing on the assessment of capstone experiences, communication, critical thinking and portfolio development, a combination of percentage agreement and a correlation coefficient was used to measure inter-rater reliability [1, 7, 9, 12, 14, 15].

As reported above, inter-rater reliability is often reported based on percentage agreement [10]. Percentage agreement, more specifically, is calculated by looking at the number of variables on which the raters gave the same score, versus the total number of variables under examination. However, a concern with this methodology is that it does not account for chance agreement (for example, agreement when raters are simply guessing), and so Cohen's Kappa, a type of correlation coefficient, was introduced to allow for this possibility and adjust the inter-rater reliability in the case of two raters [4], while the Fleiss Kappa, an adaptation of Cohen's Kappa, is used when there are three or more raters [6].

Cohen's kappa is calculated using the following formula:

$$\varkappa = Pr(a) - Pr(e) / 1 - Pr(e)$$

Where Pr(a) represents the actual observed agreement, and Pr(e) represents the chance agreement, with total

sample size calculated based on the total number of observations by the raters. A weighted version of Cohen's kappa is used to weight disagreements differently, and hence is considered to be particularly helpful for ordinal scales [5].

Cohen's kappa ranges from -1 to +1, with +1 representing "perfect agreement" between raters and -1 representing "perfect disagreement". According to Cohen, $\varkappa$ values of 0.01-0.20 indicate no to slight agreement, 0.21 - 0.40 fair agreement, 0.41 - 0.60 moderate agreement, 0.61 - 0.80 substantial agreement, and 0.81-1.00 almost perfect agreement [4].

To interpret Fleiss' $\varkappa$, we used a scale developed by Landis and Koch in which $\varkappa$ values of < 0 indicated poor agreement, 0.01-0.20 indicated slight agreement, 0.21 - 0.40 fair agreement, 0.41 - 0.60 moderate agreement, 0.61 - 0.80 substantial agreement, and 0.81-1.00 almost perfect agreement [9].

It should be noted that both Cohen and Fleiss' Kappa may lower the estimate of agreement by too great a margin, as it is difficult to estimate how much "guessing" is really taking place [10]. In other words, if there is a strong level of agreement across raters, and in particular if conditions for good inter-rater reliability have been met (competent raters, appropriate training, appropriate rubric requirements and clear descriptions of the criteria at the stated competency levels), it's possible that the strong level of agreement can be fully or mostly attributed to the rubric and the raters' interpretation of the rubric, and not to chance.

3. METHODOLOGY

All students enrolled in the thesis course were invited to a focus group to discuss the thesis experience. Twenty-four students participated, representing a variety of supervisors, program majors (and therefore thesis topics), as well as a variety of performance levels. The results of the focus group were described in [13], and students were also asked to self-assess using the rubric. These self-assessments are not discussed in this paper, due to time limitations, but will be the subject of a future paper. Supervisor data was taken from the actual, final assessments of student work. From those twenty-four, two generalist expert readers then independently scored twelve randomly chosen theses. As shown above, the thesis rubric contains eighteen criteria, grouped into 5 sections (4 with 3 categories, one with six), with four levels of achievement for each criteria; poor, adequate, good, and exceeds expectations. In total, then, we measured the agreement between raters for 216 scoring decisions, in which there were four ordinal categories.

We calculated the frequency of agreement (otherwise stated as percentage agreement), first, between the two

generalist experts; then between the two generalist experts and the student's supervisor. Since frequency of agreement does not take into account the possibility of chance agreement, we use Cohen's Kappa to measure inter-rater reliability between the two generalist readers, choosing the weighted Kappa because of the ordinal nature of the data set. To assess the three way agreement between the two generalists and the supervisor, we used Fleiss's Kappa. Finally, we assessed the frequency of agreement for each rubric criterion and group to identify which components were particularly problematic for both two way and three way agreement.

4. RESULTS AND DISCUSSION

4.1 Overall Inter-rater Reliability

The following results were generated from the analysis. Between the two expert generalist readers, we saw 47 points of disagreement, giving us the following results:

Frequency of Agreement (two generalist raters): **77.8%**
   Cohen's $\varkappa$ (unweighted) = 0.6439
   Cohen's $\varkappa$ (weighted) = 0.7079

Based on the scale developed by Cohen identified above [5], our weighted Cohen's $\varkappa$ indicates **substantial** agreement. Of note is the fact that we saw no disagreements where the two expert generalist raters differed by more than one point on the scale. This degree of agreement initially suggests that the rubric is a valid tool for improving uniformity of assessment across the projects. However, it should be acknowledged that the two expert generalists collaborated on the development of the rubric and are involved in the thesis experience in similar ways – from the perspective of supporting the development of thesis documentation and the research experience more generally.

In the three way analysis between the expert generalists and the respective supervisors, there was much more disagreement: 146 out of 216. This gives us the following results:

Frequency of Agreement (two generalist raters + supervisor): **32.41%**
   Fleiss' $\varkappa$ = 0.2577

Based on the scale developed by Landis identified above, our Fleiss's $\varkappa$ indicates **fair** agreement. This is perhaps not surprising, given that the supervisors did not design or receive formal training in using the rubric. In fact, this is one of the findings that highlights a need to better understand supervisor interpretation of the rubric. Supervisors engaged in the Engineering Science Thesis course typically see 1-2 projects, where the expert generalists see a greater range of work based on ongoing engagement with several students in the course. Exemplars, with a sample assessment rubric, may be considered to better support the supervisors.

4.2 Inter-rater Reliability by Section

Beyond traditional measures of inter-rater reliability, we also wanted to identify areas of most agreement/ disagreement both in the two and three way analyses.

Between the expert generalists, the most agreement was found in:
A. "Establishes a clear research gap/design problem, makes a convincing case for the significance of proposed research work"
B. "Describes methods or design in sufficient detail to enable understanding of work done";
C. "Document length, formatting, structure meets stated requirements, and specific demands of thesis topic."

Agreement on the third of the criteria above can be easily understood since that criterion allows for the most objective measurement: page length, adherence to formatting requirements set out by the thesis assignment statement. The agreement on the first two criteria, however, are more significant. Both of those criteria are pivotal to a general understanding of the motivation and nature of the work, and that high agreement was achieved in those two areas is promising.

Most disagreement between the generalist experts was found on the following rubric requirements:
A. "Identifies goal for thesis work that explicitly addresses this gap/problem; provides clear purpose statement"
B. "Outlines significance of research done, identifies potential future work that arises from thesis work"
C. "Abstract concisely summarizes purpose, methods, key results of research, and presents conclusions clearly"

The first of these is interesting because it is the corollary of the first point of significant agreement: even though interpretations of gap identification are consistent, these results point to differences in opinion on how explicit and detailed the goal statement should be should be. The second point of major disagreement is also understandable in that the significance and future work of the thesis may be described to varying degrees, and perhaps one expert generalist had expectations for greater depth than the other. Finally, since the third assesses an entirely stand alone section of the thesis – the abstract – the level of disagreement here is also understandable and may be attributed to disagreements in the level of detail needed.

In the three-way analysis, on the other hand, the most agreement was found in the Methods section of the rubric,

and in the criterion for visuals integration. This might correspond to the procedural nature of the methodology section, and the simpler rhetorical task of integrating visuals.

The most disagreement was found in the Discussion section, in particular with the criteria of "Engages with and explains results intelligently." Given the makeup of the readers, it is also understandable, given that the supervisor is judging the interpretation of data from a subject matter expert's point of view, while the generalist may be looking at it from a rhetorical perspective, and assessing how well have they backed the claims they are making with the available data.

These points of disagreement, for both the two- and three-way analysis, suggest points of greater intervention, either via the training of supervisors, or in the revision of the rubric elements.

4.3 Other Emergent Themes

Our analysis of these rubrics also demonstrated some challenges involved in mandating supervisor use of the provided rubrics. Several of the supervisors demonstrated little range in their scoring of the student, indicating the same level of performance across the board (eg. 4s in every category). Both the generalist experts' assessments of these particular thesis documents demonstrated significantly greater range, suggesting that some supervisors may not be assessing carefully according to the criteria provided.

The second, other noticeable finding relates to the correspondence between the rubric's description of levels of achievement and the grade awarded to the student. In the rubric, the four levels of achievement correspond to various grade ranges: 1 - Fails (0-60%); 2 - Adequate (60-70%); 3 - Good (70-80%); 4 - Exceeds expectations (80-100%). In most cases, awarded grades corresponded well to the categories most often checked by supervisors; students scoring the majority of components in category 4, for example, were awarded with grades above the 80% mark. Students with a mix of scores in the 3-4 range received in the 70-85% range. While supervisors were internally consistent in their rubric and numerical assessment of students, significant disagreement between the generalist experts and the supervisor on the overall quality of the thesis document was recorded on several occasions. In most cases, the supervisor assessment was more generous than that of the generalist experts, except for those theses considered to be exceptional.

It should also be noted that there are limitations to the capabilities of the generalist experts. While the generalist experts are very familiar with the thesis course, the requirements of a strong research project and the course deliverables, there are dimensions of the thesis that they are not familiar with: for example, did the student describe all of their methods/results appropriately? How do we know whether the literature review represents state of the art in the field? Is the future work outlined appropriate, given the state of the field and the results of the thesis?

4. CONCLUSIONS AND FUTURE WORK

Our work on rubrics in this course, via studies in rubric design, student focus groups, and other means, has been geared at assessing the rubric's validity as a multi-purpose tool. Our analysis of the inter-rater reliability of the rubric was intended to help assess its validity as a means of developing consistency in the scoring of students. Our findings suggest that there is substantial agreement between raters who are highly familiar with the rubric, but only fair agreement across those raters and the supervisor. A deeper understanding of the scoring approach of the thesis supervisors would be useful, particularly their thought process when completing the rubrics. Focus groups or interviews with thesis supervisors will be the next step in this work.

References

[1] M.J. Bresciani, M. Oakleaf, F. Kolkhurst, C. Nebeker, J. Barlow, K. Duncan, and J. Hickmott, "Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines," in *Practical Assessment, Research & Evaluation,* vol. 14, no. 12, 2009.

[2] A. Chong and L. Romkey, "Adapting existing assessment tools for use in assessing engineering graduate attributes", in *Proc. CEEA Canadian Engineering Education Conf., CEEC12,* Ron Britton (ed.) (Winnipeg, MB, 17-20 June 2012), 16 pp., 2012.

[3] A. Chong and L. Romkey, "Pedagogical innovations to support large scale undergraduate independent research projects", in *Proc. CEEA Canadian Engineering Education Association Conf., CEEC14*, Sean Maw (ed.) (Canmore, AB; 8-11 June 2014), 1 pp., 2014.

[4] J.A. Cohen, "A coefficient of agreement for nominal scales", *Educ Psychol Meas*, vol. 20, pp. 37-46, 1960.

[5] J.A. Cohen, "Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit", *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[6] J.L. Fleiss, "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[7] R. Gerlick, D. Davis, M. Trevisan and S. Brown, "Establishing inter-rater agreement for TIDEE's teamwork and professional development assessments", in *Proc. ASEE American Society for Engineering Education Conf., ASEE11,* (Vancouver, BC, 26-30 June 2011), 14 pp., 2011.

[8] A. Jonsson and G. Svingby, "The use of scoring rubrics: Reliability, validity and educational consequences", *Educational Research Review,* vol. 2, pp. 130-144, 2007.

[9] J.R. Landis and G.G. Koch, "The measurement of observer agreement for categorical data", *Biometrics*, vol. 33, pp. 159-74, 1977.

[10] M. McHugh, "Inter-rater reliability: the kappa statistic", *Biochemia Medica*, vol. 22, no. 3, pp 276-82, 2012.

[11] B.M. Moskal and J.A. Leydens, "Scoring rubric development: validity and reliability", *Practical Assessment, Research and Evaluation,* vol. 7, no. 10, 2000.

[12] J. Newell, K. Dahm, and H. Newell, "Rubric development and inter-rater reliability issues in assessing learning outcomes", in *Proc. ASEE American Society for Engineering Education Conf., ASEE02*, (Montreal, QC, 16-19 June 2002), 8 pp., 2002.

[13] L. Romkey, A. Chong, and L. El Gammal, "Using student focus groups to support the validation of rubrics for large scale undergraduate independent research projects", in *Proc. CEEA Canadian Engineering Education Association Conf., CEEC15, Thomas Doyle (ed.)* (Hamilton, ON; 31 May – 3 June 2015), 8 pp., 2015.

[14] E. Saxton, S. Belanger and W. Becker, "The critical thinking analytic rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments." *Assessing Writing,* vol. 17, pp. 251-270, 2012.

[15] T. Utschig, J. Norback and J. Bryan, "Workforce communication instruction: preliminary inter-rater reliability data for an executive-based oral communication rubric", in *Proc. ASEE American Society for Engineering Education Conf., ASEE12* (San Antonio, TX; 10-13 June 2012), 15 pp., 2012.

APPENDIX A: RUBRIC AND RUBRIC GUIDE

FINAL THESIS REPORT RUBRIC – I – ESC499

Student Name: _____ Supervisor: _____ Grade: ____ /100

Component	1	2	3	4	Requirement	Comments (*Use back if necessary*)
Introduction	☐	☐	☐	☐	Establishes context necessary to facilitate thorough understanding of thesis work in a concise manner	
	☐	☐	☐	☐	Establishes a clear research gap/design problem, makes a convincing case for the significance of proposed research work	
	☐	☐	☐	☐	Identifies goal for thesis work that explicitly addresses this gap/problem; provides clear purpose statement	
Literature Review / Background	☐	☐	☐	☐	Explains theoretical concepts important to understanding of thesis work	
	☐	☐	☐	☐	Identifies, summarizes, and synthesizes relevant research in constructing an understanding of current state of field	
	☐	☐	☐	☐	Enables deeper understanding of research question/design problem through analysis of research in the field, indicating a path for moving research forward	
Methods and Findings	☐	☐	☐	☐	Describes methods or design in sufficient detail to enable understanding of work done	
	☐	☐	☐	☐	Provides justification for methods chosen or design decisions made	
	☐	☐	☐	☐	Results displayed clearly in organized manner, using appropriate figures or graphics; key results highlighted	
Discussion and Conclusions	☐	☐	☐	☐	Engages with and explains results intelligently	
	☐	☐	☐	☐	Identifies key claims to be drawn from results of research or design evaluation, qualifies them appropriately	
	☐	☐	☐	☐	Outlines significance of research done, identifies potential future work that arises from thesis work	
Overall Document Design:	☐	☐	☐	☐	Abstract concisely summarizes purpose, methods, key results of research, and presents conclusions clearly	
	☐	☐	☐	☐	Document length, formatting, structure meets stated requirements, and specific demands of thesis topic	
	☐	☐	☐	☐	Organized well, with content in discrete and appropriate positions in paper, structure clearly laid out, transitions that create flow in document	
	☐	☐	☐	☐	Demonstrates grammatically correct, coherent prose that concisely and clearly communicates complex topics in well designed paragraphs and sentences	
	☐	☐	☐	☐	Uses and integrates well-designed visuals effectively to communicate key concepts / results	
	☐	☐	☐	☐	Provides clear attribution of ideas throughout paper using a known referencing standard; uses references effectively to help establish context, back claims, or justify decisions	
Project Experience	☐	☐	☐	☐	Work has contributed to scholarship in field / made a measurable impact	
	☐	☐	☐	☐	Demonstrated initiative, ability to work independently, time management skills and ownership of work throughout thesis project	
	☐	☐	☐	☐	Thesis work posed a significant challenge, requiring superb engineering & scientific knowledge and skills	
	☐	☐	☐	☐	Quality of effort and thesis work indicative of potential for future research success	
	☐	☐	☐	☐	Has incorporated feedback and additional research on initial deliverables to improve final thesis document and work	

*1 - Fails (0-60%); 2 - Adequate (60-70%); 3 - Good (70-80%); 4 - Exceeds expectations (80-100%). These numerical equivalents are only approximate; final grade and value of each component is up to the supervisor.

A ROUGH GUIDE TO THE FINAL THESIS REPORT RUBRIC GUIDE – I – ESC499

Component	Fails	Adequate	Good	Exceeds Expectations
Introduction	Missing key elements of context necessary to understand thesis work	Establishes just sufficient context necessary to facilitate a basic understanding of thesis work	Establishes context necessary to facilitate understanding of thesis work	Develops context appropriately and concisely in facilitating thorough understanding of thesis work
	Research gap or design problem remains unarticulated or unclear	Research gap or design problem is identified, but too broad or general to define project clearly	Identifies a clear research gap/design problem	Establishes a clear research gap or design problem, makes a convincing case for the significance of proposed research work
	Goal of thesis work is difficult to identify, or unrelated to gap or problem statement	Stated goal for thesis work is vague, imprecise, or not clearly related to gap/problem statement	Identifies goal for thesis work that addresses this gap/problem	Explicitly identifies goal for thesis work in a clear purpose statement for the project that addresses gap/problem
Literature Review / Background	Missing key explanations of theoretical concepts important to thesis work	Most theoretical concepts important to work are identified and briefly explained	Identifies and explains theoretical concepts important to understanding of thesis work	Explains theoretical concepts clearly, concisely, in context of thesis work
	Fails to acknowledge or reference key research/prior work in the field	Identifies some important research/prior work in the field, but misses a few essential developments	Identifies and summarizes most of the key research/prior work in developing an understanding of the field	Identifies, summarizes, and synthesizes relevant research in constructing a nearly complete understanding of current state of field
	Analysis of field is incomplete and fails to further develop the research gap/design problem	Analysis of field provides limited help to further develop the research gap/design problem	Enables understanding of research question/design problem through analysis of research in the field	Enables deep understanding of research question/design prob. via thorough analysis of research in the field, indicating path for moving forward
Methods and Findings	Fails to explain key elements of methods or design	Most elements of methods or design are explained in sufficient detail; missing elements may hamper understanding of work done	Describes methods or design in sufficient detail to enable understanding of work done	Detailed description of methods or design helps facilitate a thorough understanding of project
	Fails to justify key elements of method or design decisions	Provides some justification for most methods chosen/design decisions made	Provides sufficient justification for methods chosen / design decisions made	Justification for methods chosen/design decisions clearly and convincingly articulated, warranting validity of project
	Results not displayed in organized or appropriate manner	Most results are displayed in an organized manner, using some appropriate figures or graphics	Results displayed clearly in organized manner, using appropriate figures or graphics	Results displayed clearly in organized manner, using appropriate visuals that help highlight key results and findings
Discussion and Conclusions	Engages with results only superficially, without explanation of significance	Engages with and provides explanation for most results	Engages with and explains key results intelligently	Employs and explain results clearly in the context of research / design claims made
	Fails to make key claims from results of research or design evaluation	Makes appropriate claims from results of research or design evaluation, though claims may not be fully warranted	Identifies and explains key claims to be drawn from results of research or design evaluation	Identifies and warrants key claims to be drawn from results of research or design evaluation, qualifies them appropriately
	Fails to identify significance of research or design work done	Summarizes research / design work done, but fails to place it in context of prior or future work	Clearly identifies significance of research/ design work done in context of past work	Outlines significance of research/design work done, identifies potential future work that arises from thesis work
Overall Document Design:	Abstract fails to adequately describe nature and conclusions of project	Abstract provides a vague description of nature and conclusions of project	Abstract summarizes key elements of thesis sufficiently	Abstract concisely and completely summarizes purpose, methods, key results of research, presenting conclusions clearly
	Document length or formatting fails to meet many of the stated requirements	Some inconsistencies in formatting, but mostly meets the stated requirements	Document length, formatting, structure meets stated requirements	Document length, formatting, structure meets stated requirements, and specific demands of thesis topic

*1 - Fails (0-60%); 2 - Adequate (60-70%); 3 - Good (70-80%); 4 - Exceeds expectations (80-100%). These numerical equivalents are only approximate; final grade and value of each component is up to the supervisor.

A ROUGH GUIDE TO THE FINAL THESIS REPORT RUBRIC GUIDE – I – ESC499

Component	Fails	Adequate	Good	Exceeds Expectations
	Poorly organized throughout, lacking clear structure and flow throughout document	A few organizational problems detract from an otherwise well structured thesis	Organized well, with content in discrete and appropriate positions in paper, and structure clearly laid out	Organized well, with logical and explicit structure, and transitions that create flow throughout document
	Grammatical problems hamper understanding of key elements of the document	Some grammatical problems hamper readability of prose, but not understanding of document	Demonstrates grammatically correct, coherent prose throughout	Demonstrates grammatically correct, coherent prose that concisely and clearly communicates complex topics in well designed paragraphs and sentences
	Visuals not used appropriately or well integrated into document	Visuals employed appropriately, but may not be well integrated into document	Uses and integrates visuals effectively to communicate key concepts / results	Uses and integrates well-designed visuals effectively to communicate key concepts / results
	Poor attribution of ideas throughout, missing key references and failing to use an appropriate referencing standard	Provides clear attribution of ideas throughout paper using a known referencing standard, with only a few gaps or errors	Provides clear and thorough attribution of ideas throughout paper using a known referencing standard	Provides clear attribution of ideas throughout paper using a known referencing standard; references used effectively to help establish content, back claims, or justify decisions
Project Experience	Work makes no contribution to the field, and is only marginally relevant to current good scholarship	While no significant impact is made, work is relevant to good scholarship in the field	Work advances the possibility for future advancements in the field	Work has contributed to scholarship in field / made a measurable impact
	Student lacks the ability to take initiative, work independently, manage their time and/or take ownership over the project, negatively impacting the thesis experience	Student sometimes shows the ability to take initiative, work independently, manage their time and take ownership over the project	Student takes initiative and works independently, demonstrating good time management skills and project ownership	Demonstrated excellence in taking initiative, ability to work independently, time management skills and ownership of work throughout thesis project
	Thesis work completed did not meet expectations, nor showcase the knowledge and skills expected from an undergraduate engineering student	Thesis work required the knowledge and skills expected of an undergraduate engineering student	Thesis work posed a reasonable challenge, requiring engineering & scientific knowledge typical of a senior undergraduate or junior graduate student	Thesis work posed a significant challenge, requiring superb engineering & scientific knowledge and skills
	Student does not demonstrate potential for future research work	Quality of effort and thesis work indicative of some potential for future research work, with more study and experience	Quality of effort and thesis work indicative of some potential for future research success	Quality of effort and thesis work indicative of strong potential for future research success
	Student does not incorporate feedback or additional research from initial deliverables in the final thesis document or project work	Student incorporates some feedback and/or additional research from initial deliverables in the final thesis document or project work	Student incorporates most feedback and/or additional research from initial deliverables in the final thesis document or project work	Student has incorporated all relevant feedback from initial deliverables, and has actively incorporated new research or analysis to improve final thesis document and project work

*1 - Fails (0-60%); 2 - Adequate (60-70%); 3 - Good (70-80%); 4 - Exceeds expetitions (80-100%). These numerical equivalents are only approximate; final grade and value of each component is up to the supervisor.