Routledge
Taylor & Francis Group

ASSESSMENT & EVALUATION in HIGHER EDUCATION
Volume 46 Number 8 December 2021

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Examiners' use of rubric criteria for grading bachelor theses

Marjolein Haagsman, Basten Snoek, Anton Peeters, Karin Scager, Frans Prins & Martijn van Zanten

To cite this article: Marjolein Haagsman, Basten Snoek, Anton Peeters, Karin Scager, Frans Prins & Martijn van Zanten (2021) Examiners' use of rubric criteria for grading bachelor theses, Assessment & Evaluation in Higher Education, 46:8, 1269-1284, DOI: 10.1080/02602938.2020.1864287

To link to this article: https://doi.org/10.1080/02602938.2020.1864287

Published online: 19 Jan 2021.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
2021, VOL. 46, NO. 8, 1270–1285
https://doi.org/10.1080/02602938.2020.1864287

Routledge
Taylor & Francis Group

Examiners' use of rubric criteria for grading bachelor theses

Marjolein Haagsman, Basten Snoek, Anton Peeters, Karin Scager, Frans Prins and Martijn van Zanten

Utrecht University, Utrecht, The Netherlands

ABSTRACT

Students are generally required to demonstrate diverse skills when writing their bachelor thesis. Accordingly, examiners are expected to consider all these skills when assessing the thesis, regularly with one overall grade. In this study, we examine which criteria of a rubric contribute most to the overall assessment. The study is performed through quantitative analyses of 318 theses of undergraduate biology students. The analyses demonstrate that all criteria scores are predictive, but that *scientific quality* and *professional attitude* give the best prediction of thesis grade, together with *structure*. The predictiveness of *scientific quality* and *professional attitude* correspond with the instructions given to examiners that these are important criteria to consider. Presentation-related criteria scores on *writing skills* and *expressing catchy and justifying titles* give the lowest prediction of grade. This study identifies that some criteria appear more predictive for low grades than for high grades, with *professional attitude* being a good predictor for low grades and *abstract* being a good predictor for high grades. We recommend similar analyses for students to help them prioritise the most relevant criteria, for supervisors to instruct students on these criteria, and for education managers to evaluate whether bachelor theses are assessed on the criteria they find most relevant.

KEYWORDS

Rubrics; summative assessment; bachelor's thesis; scientific writing

Introduction

Undergraduate students in the Netherlands commonly complete their bachelor studies with a 'capstone' thesis. This bachelor graduation thesis is considered an important part of the curriculum as it requires students to find relevant literature, process complex information, use scientific reasoning, think critically and work individually. All these skills need to be considered by the examiner when assessing the overall thesis quality and students' professional attitude, often with only one single grade. Thus, the examiner is required to interpret the quality of each assessment criterion and then needs to integrate all criteria to produce an overall quality score or grade. The current study explores how examiners weigh different rubric criteria in the overall assessment of undergraduate science theses. We quantified and analysed assessment criteria for the biology undergraduate programme of Utrecht University. We aim to examine which criteria are considered most significant for the overall grade of research literature theses.

CONTACT Marjolein Haagsman m.e.haagsman@uu.nl

Quality criteria of bachelor theses

The bachelor thesis used for the current study consists of a review of existent primary research literature in the field of biology in the broadest sense. The main criteria for writing a good bachelor thesis in this study are to: i) formulate a testable question, ii) select and process information from literature, iii) critically evaluate literature, and iv) write a concise scientific report. The steps of the research cycle are generally reflected in the required components of the thesis such as introduction, results and discussion (Prins, de Kleijn, and van Tartwijk 2017). In the field of biology, Timmerman et al. (2011) developed a set of generic criteria to assess the components needed in the process of scientific writing. Criteria included were context and accuracy of the introduction, scientific merit and testability of the hypothesis, experimental design, data selection and presentation, analysis of results, significance and final conclusions of the discussion, alternative explanations and limitations of the discussion, use of primary literature and overall writing quality (Timmerman et al. 2011). Although these criteria were developed to assess reports in undergraduate biology laboratory courses, most of them are applicable to every scientific writing assignment and are used in this study.

Reddy and Andrade (2010) observed a gap in higher education research and endorsed more studies on the correlation of the quality of these criteria and final grading. To our knowledge, only a few studies have explored which criteria are of prime relevance when grading scientific writing assignments or bachelor theses. One study shows that thesis grades are mainly related to the assessment of the research set-up and interpretation of the results (Bourke and Holbrook 2013). These criteria are however irrelevant for bachelors' theses that are solely based on literature research and do not require students to design a research set-up and obtain data. Another study on bachelors' theses in nursing found that thesis grades mainly correlate with student's ability to compare the reported studies (Lundgren, Halvarsson, and Robertsson 2008). The theses described by Lundgren, Halvarsson and Robertsson are, however, not assessed with rubrics describing quality levels for every criterion. Here, we specifically assess the relationship of bachelor thesis grades and criteria scores that are acquired with the aid of such rubrics.

Goals of using rubrics from the student and examiners' perspective

Rubrics gained popularity in higher education over the last thirty years and are commonly referred to as '*a document that articulates the expectations for an assignment by listing the criteria, or what counts, and describing levels of quality from excellent to poor*' (Reddy and Andrade 2010, 1). In other words, rubrics define criteria of a certain task and/or expectation and contain typifying descriptions of various quality levels for each of these criteria. The examiner uses the rubric to address the quality levels that fits best with the students' work. If well designed, rubrics help students to assess their own work and direct them in what to improve for upcoming drafts or tasks (Tai et al. 2018). Rubrics are designed to make students aware of the expectations of the examiners, which likely explains why they have been shown to reduce students' anxiety towards assignments (Andrade and Du 2005; Panadero, Alonso-Tapia, and Reche 2013). On the other hand, rubrics guide teachers and examiners in providing relevant feedback (Prins, de Kleijn, and van Tartwijk 2017). From a governance viewpoint, rubrics can be administered alongside the graded final thesis, which facilitates (external) committees like the board of examiners and accreditation panels in validating the quality of the product, and curriculum as a whole. In this study we focus on how examiners use rubrics for assigning grades for summative assessment.

Rubrics for summative assessment

The summative assessment of theses is a complex process since many criteria are involved (Sadler 2009). There is not one single answer or best approach for writing such intricate

assignments. In this study, examiners are provided with a rubric to guide them in the complex process of assessing bachelor graduation theses. The examiners can use the rubric to review and assess the distinct criteria, but no formula is provided to calculate final grades from these criteria scores. Thus, the examiners are free in how to use the rubric for determining the final overall grade.

It is logically relevant that these weighed grades accurately reflect the quality of students' work. Students in higher education greatly value the grades they obtain, as they do not perceive them as merely feedback instruments, but also as measurements of academic achievement needed for entrance to MSc programmes and even their future career (Goulden and Griffin 1995). This is especially true for bachelor theses. Grades are however influenced by the experience and skills of the examiner, even when using rubrics with well-defined criteria (Kapborg and Berterö 2002). Different examiners might put different 'weights' to each rubric criterion (Sadler 2009). Hence, it is mainly the examiner that determines the validity of a grade and not the instrument itself (van der Vleuten et al. 2012). This is especially an issue for the bachelor theses within this study, since these are assessed by a large variety of examiners from PhD students to full professors, and from both within and outside the university and across biological disciplines, from ethics to molecular cell biology and from marine ecology to bioinformatics. Thus, we aim to improve insight into how thesis grades are assigned by examiners.

The rubric used in this study contains categories based on commonly required components and criteria of science reports, including; *title*, *summary*, *introduction, set-up*, *discussion and conclusion*, *scientific quality*, *spelling and grammar*, *style, length and layout*, *figures and tables*, *references* and *professional attitude*. Each criterion includes descriptions of subcriteria relevant for literature studies and comparable to the sub-criteria described by Timmerman et al. (2011). Here, we explore how examiners weigh these criteria to decide upon the overall thesis grade. More specifically, we aim to address the following questions:

1. What is the relation between criteria scores and thesis grade?
2. Which criteria are most predictive for grades of bachelor theses?

This study is performed through quantitative analyses of criteria scores and thesis grades of 318 bachelor theses of three subsequent years in our Utrecht University biology curriculum.

Materials and methods

Participants

The rubrics included in our dataset were used to assess 318 students who wrote their thesis within the broad biology bachelor programme of Utrecht University. Rubric scores are only documented after students pass the course (thesis grade of 5.5 or above, scale 1.0 to 10.0). The theses evaluated were from all students who started their thesis project in the academic years 2014–2015, 2015–2016 and 2016–2017 (129, 138 and 51 students). Of these, 156 (49.1%) were written by males and 161 (50.9%) by females. Henceforth, only the rubric assessments and thesis grades were collected and anonymously included in this study. The rubric assessments and thesis grades are always agreed on by two independent examiners, of which at least one is an Utrecht University biology faculty member or approved faculty member of an affiliated department within the university and certified with at least a basic teaching qualification. In total, 202 individual examiners were involved in grading. Of all theses, 12.3% were graded by external supervisors from companies, institutes or non-governmental organizations and had no direct affiliation with Utrecht University. Of all theses, 0.2% were graded by a technician, 5.2% by a PhD student, 2.7% by 'teachers', 45.3% by assistant professors, 12.5% by associate professors and 17.1% by full professors.

Context

It is compulsory for every Utrecht University biology bachelor student to write a theoretical bachelor thesis that addresses a scientific question that falls within one of the many biological disciplines, including associated disciplines such as didactics, ethics, (sustainability) policy and consultation. The thesis should contain a well-defined research question that can be answered using mostly primary peer-reviewed scientific literature. The product is a well thought-over text of about 6000–8000 words and may be written in English or in Dutch, with at least 20 primary sources cited but usually many more. Next to the compulsory thesis, the majority of students also conducted a research project in the field or laboratory on the same or a similar topic.

Course setting

Students spend 10 weeks half-time or 5 weeks full-time on the writing process (total 200 h). There are three compulsory seminars. In the first week an introduction lecture is organised where rules, guidelines and expectations are explained. During this session, the students are advised to look at an online version of the rubric to get a view on the expectations of the examiners (see Appendix 1). In a second meeting organised by the Utrecht University Library, the students practice with literature searches, correct literature usage and referencing. Lastly, in week 3 of the course, aspects of academic writing are trained. For the rest of the time, students work individually under direct guidance of their topical (daily) supervisor. It is compulsory to hand-in a thesis plan (raw version of main question and ideas), on a set deadline date (approximately week 3).

During the writing process, the student usually hand-in two versions of their thesis: a concept version and a final version. The student receives feedback on both versions from the daily supervisor on the content, structure, effort and the progress. Examiners assess the final version of the thesis with one grade that considers all assessment criteria.

Rubric scores and grading

The final version of the thesis is assessed with the compulsory aid of a rubric. The rubric was developed based on examples from the literature by a panel of experienced examiners. The rubric was used in the following year to assess theses and was evaluated thereafter. Based on this evaluation, adjustments were made that led to the rubric that is evaluated in this study (see Appendix 1). This rubric describes criteria for 13 categories typical for theses in the science domain (*title, abstract, introduction, scope, structure, discussion, conclusions, scientific quality, spelling and grammar, style, length and lay-out, figures, references*) and the writing process (*professional attitude*). Each category is further divided into 1–6 sub-criteria. Examiners can tick boxes for each of these sub-criteria to judge as *insufficient, sufficient* or *good*. It is allowed to tick boxes in two sub-criteria, if the supervisor deems the level of the student is between *insufficient/sufficient* or between *sufficient/good*.

Students are graded from 1.0 (lowest) to 10.0 (highest). If the thesis is marked a 5.5 or higher the students receive 7.5 European Credits (ECs). If the examiners grade the thesis as insufficient ($<$5.5), the student needs to improve the work and is allowed to hand-in a much-revised version. If the examiners deem that it is unlikely that the required level is met by modifying the work, the student needs to start over with the thesis course and must find a different supervisor and examiner.

The rubric is purposely designed as an evaluation guideline for the examiners and as feedback tool for the student. The examiners assign one overall judgment (grade) based upon the rubric criteria. However, the criteria *scientific quality* and *professional attitude* are highlighted to advise examiners to particularly consider these two criteria when deciding upon the grade. Although explicitly not designed as a calculation table, if all criteria of the thesis are assessed at the '*sufficient*' level, the suggested grade is 7.

Quantitative assessment of rubric scores

To allow quantitative evaluation required for this study, we numerated all sub-criteria. A 1.0 was assigned if the sub-criterion was judged as *insufficient*, a 2.0 when *sufficient* and a 3.0 if deemed *good*. If two boxes for one sub-criterion were ticked the average was calculated (i.e. 1.5 or 2.5). A 0 (null) was assigned when none of the boxes of a particular category was ticked. The subcriteria within the *introduction and scope* category could not be directly numerated, since the number of tick boxes per assessment level differed; i.e. the *insufficient* category contains two boxes for *introduction* and two for *scope*, while respectively four and two sub-criteria describe the *introduction* at the *sufficient* and *good* assessment levels, and one at each level the *scope*. To circumvent confounding effects, we therefore assigned a 1.0, 2.0 or 3.0 to the *introduction* and *scope* categories as a whole, without taking the subcriteria into account. A 1.0 was for instance assigned if only box(es) at the *insufficient* assessment level was/were ticked, a 2.0 if only boxes at the *sufficient* level was/were ticked and a 3.0 if only boxes at the *good* level was/were ticked. A 1.5 or 2.5 were assigned if boxes in more than one category was ticked, even if for instance four boxes of the *introduction* category were ticked at the *sufficient* assessment level and only one at the *insufficient* level. A similar method was applied to the *title* category and the conclusion part of the *conclusions and discussion* category, that both contained two sub-criteria at the *insufficient* level and only one at the *sufficient* and *good* assessment levels. Besides the numeration of individual sub-criteria, we calculated per rubric which (sub)categories were ticked at two assessment levels (i.e. between *insufficient* and *sufficient* or between *sufficient* and *good*. In addition, we calculated the average score per category.

Method of analysis

All criteria scores and grades were collected in a data matrix and analysed in R (http://www.R-project.org/), using the Openxlsx() function from the Openxlsx package (https://github.com/ycphs/openxlsx). Cronbach's alpha was determined using the alpha() function from the Psych package (Revelle 2019). Means and standard deviations (SDs) were calculated using base-R functions mean() and sd(). For linear regression of criteria scores explaining the final grade, the summary() and lm() functions of base-R were used. To estimate the relative importance of criteria scores, a linear model and the boot.relimp() and booteval.relimp() functions from the relaimpo package (Grömping 2006) were used. For visualization the packages: Ggplot2 (Wickham 2016), Cowplot (Wilke 2019) Hplots (Warnes et al. 2019), Hgdendro (de Vries and Ripley 2016) and Patchwork (Pedersen 2019) were used. The heatmap.2() function from the Gplots package was used to calculate the dendrogram which was then extracted and used for visualisation using ggplot2 and patchwork.

Results

Descriptive statistics of criteria scores and thesis grades

The 318 bachelor theses in this study were assessed with an average grade of 7.58 (SD = 0.80), ranging from 5.50 to 10.00. The rubrics used to guide assessment contained one or more sub-criteria per criterion that could be scored with either a 1, 2 or 3 (*insufficient, sufficient* and *good*). The sub-criterion scores were generally most similar to sub-criterion scores belonging to the same criteria (see Figure 1), suggesting that criteria stand alone and no splitting or merging of categories is needed for further analysis of the criteria scores. The criteria scores had high reliabilities for the final grade, Cronbach's $\alpha = 0.96$.

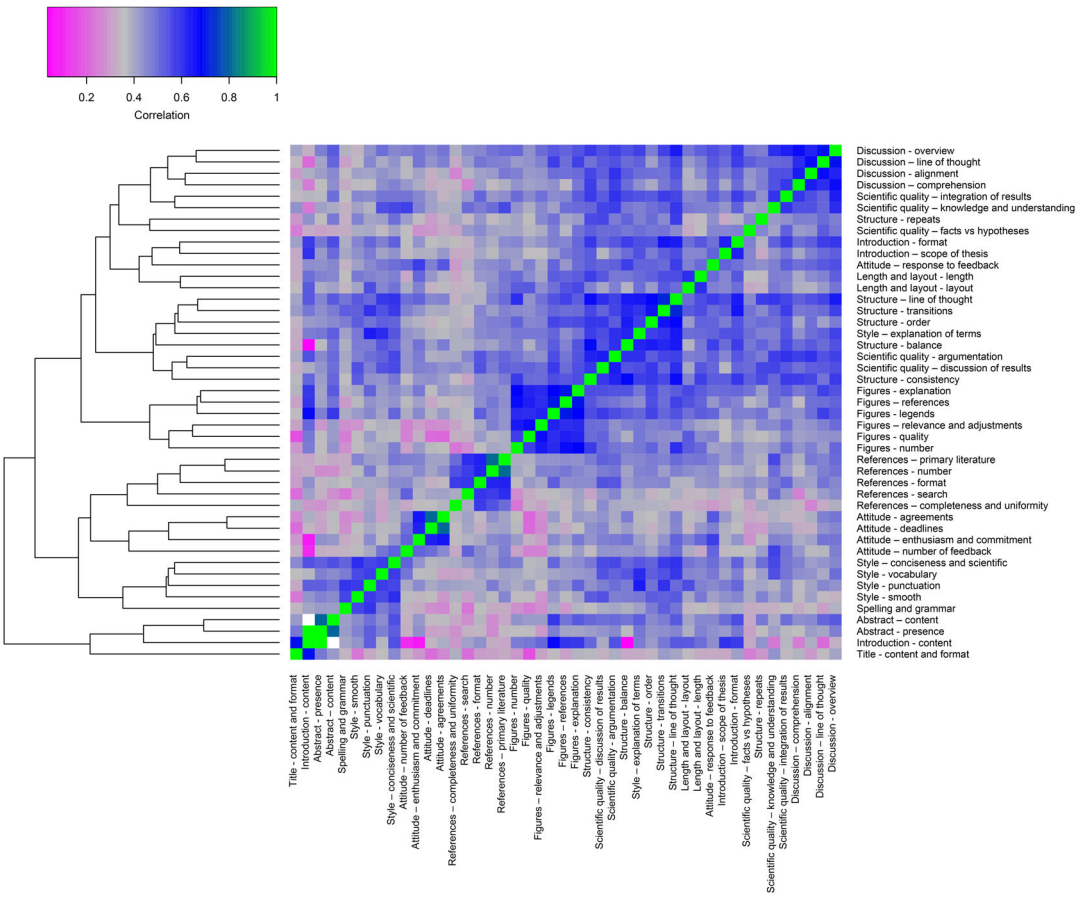

Figure 1. Calculated Euclidean distance hierarchically clustering matrix, indicating correlations between the scores of all subcriteria of all 13 categories of all numerated rubrics. The color range in the legend indicates the Pearson correlations between the subcategories (pink = Pearson correlation of 0.0, green = Pearson correlation of 1.0). The subcriteria are ordered based on the similarity in scores, also indicated by the dendogram on the left. Hence, subcriteria that are scored similarly, such as the discussion overview and discussion line of thought, are directly connected in this dendogram. The dendogram is calculated based on Euclidean distance hierarchically clustering.

Regression of criteria scores for all thesis grades

To investigate the relation between rubric criteria scores and thesis grade, we plotted the average scores of all criteria (rubric score) against thesis grade (see Figure 2). A positive relation is observed as expected, although the curve is sigmoidal. Nevertheless, rubric scores vary largely per specific thesis grade (see Figure 2). A possible explanation for the variations is that some criteria contribute more than others. Hence, we determined how each of the separate criteria scores are related to thesis grade, while correcting for the interdependence of the criteria scores (see Table 1). All criteria scores are significantly related to final grade. Since the *scientific quality and professional attitude* were marked in the rubric as 'important', it was hypothesized that these criteria had highest effects on the thesis grade. The *scientific quality* and *professional attitude* are, together with *structure*, indeed the best predictors of grade and explain respectively 62%, 53% and 58% of the variance in grades ($R^2_{adj} = 0.62$, $R^2_{adj} = 0.53$ and $R^2_{adj} = 0.58$). *Title* and *spelling and grammar* scores give the lowest prediction of grade and explain respectively 23% and 24% of the variance in grades. Multiple regression analysis of criteria on thesis grade yielded similar results (see Figure 3).

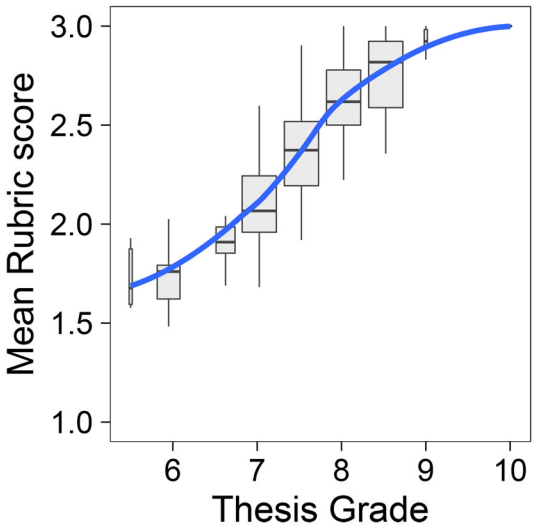

Figure 2. Trendline (blue) of the correlation between thesis grade and the mean score of the 13 subcriteria of the assessed rubric. Boxes behind the trendline indicate the median and variation in rubric score per grade. Shown are the boundaries of the second and third quartile of the data distribution. Black bars within the boxes indicate the median and whiskers the Q1 and Q4 values within 1.5 times the interquartile range. Individual dots are outliers beyond the Q1 and Q4 intervals.

Table 1. Descriptive statistics for criteria scores of all rubrics used in the assessment of bachelor theses in biology.

Rubric criterium	N[a]	Range	Mean	S.D.	r	R^2_{adj} [b]
Title	296	1.0–3.0	2.27	0.48	0.48	0.23**
Abstract	294	1.0–3.0	2.29	0.56	0.59	0.34**
Introduction	296	1.0–3.0	2.43	0.52	0.61	0.37**
Scope	281	1.0–3.0	2.32	0.60	0.56	0.32**
Structure	296	1.0–3.0	2.39	0.50	0.76	0.58**
Discussion	291	1.0–3.0	2.31	0.49	0.72	0.52**
Scientific quality	295	1.0–3.0	2.44	0.44	0.79	0.62**
Spelling and grammar	290	1.0–3.0	2.49	0.53	0.50	0.24**
Style	296	1.0–3.0	2.39	0.46	0.72	0.52**
Length and layout	293	1.0–3.0	2.42	0.51	0.59	0.34**
Figures	296	1.0–3.0	2.36	0.53	0.62	0.38**
References	296	1.5–3.0	2.58	0.43	0.61	0.37**
Attitude	287	1.0–3.0	2.52	0.46	0.73	0.53**

[a]The number of category scores differs per rubric category because of missing data.
[b]*$p < 0.01$; **$p < 0.001$.

Regression of criteria scores for sufficient, good and excellent thesis grades

Although informative, overall correlations provide limited insight into which categories contribute most to sufficient, good and excellent grades. We therefore assessed whether some rubric categories are of lower or higher relevance to examiners when assessing students with a low or high grade. Therefore, we divided all theses in three groups (bins) based on the grade: theses with grades up to and including 7.0 (sufficient), above 7.0 and up to and including 8.0 (good) and above 8.0 (excellent). We then again performed a regression analysis on all categories, but now for these three bins separately (see Table 2).

The regression analysis shows that *scientific quality* ranks among the best predictors for thesis grades in all three bins. Nonetheless, some categories are better predictors for grades in the sufficient bin than for those in the good or excellent bins. For example, theses in the sufficient bin are best distinguished by *scientific quality* ($R^2_{adj} = 0.30$) and *professional attitude* ($R^2_{adj} = 0.27$) scores. However, theses that received grades in the good bin are mainly distinguished by

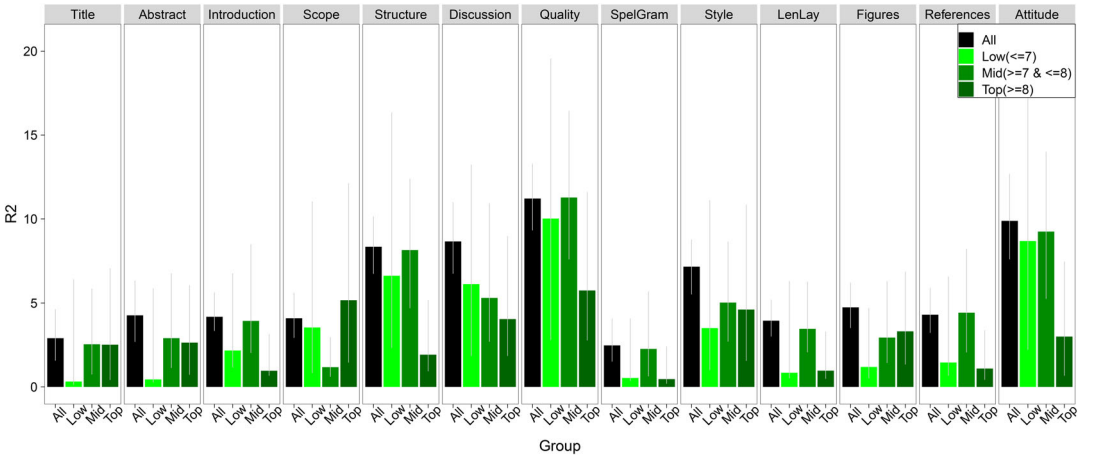

Figure 3. Multiple linear regression analysis of thesis grade and criteria scores for all theses and for theses graded as sufficient ($\leq$7.0), good ($<$7.0 and $\leq$8.0), and excellent ($>$8.0). The grey bars indicate standard deviations.

scientific quality ($R^2_{adj} = 0.25$) and *discussion* ($R^2_{adj} = 0.29$), whereas theses in the excellent bin are mainly distinguished by *scientific quality* (Adj. $R^2 = 0.13$), *abstract* ($R^2_{adj} = 0.13$) and *style* ($R^2_{adj} = 0.13$). Interestingly, *abstract* is one of the most predictive criteria in the excellent bin ($R^2_{adj} = 0.13$), but one of the least predictive criteria of the sufficient bin ($R^2_{adj} = 0.01$). Similarly, *professional attitude* is one of the most significant criteria ($R^2_{adj} = 0.27$) for the sufficient bin, although being least significant for grades belonging to the excellent bin ($R^2_{adj} = 0.00$). Thus, the regression analysis shows that some criteria appear more relevant in the assessment of theses judged as sufficient than those that received an assessment that qualifies as good or excellent.

Visualisation of criteria scores

The differences in the explained variances of criteria scores per grade bin suggest that some criteria appear of significance when assessing a sufficient, good or excellent thesis, or that some criteria are considered more determinative for an examiner to assign a grade that falls in a particular bin. We hence visualised the criteria scores per bin (see Figure 4), and of all theses separately (see Figure 5), to better understand the criteria scores provided. In Figure 4, the number of theses with a specific criteria score are shown for every bin. In Figure 5, all 318 theses are sorted in columns from the lowest thesis grade (left) to the highest grade (right). Rows represent colour-coded criteria scores with red bars for scores of 1.0, yellow bars for scores of 2.0 and green bars for scores of 3.0. For example, the thesis with the lowest thesis grade (first column on the left) was scored with a 1.43 for attitude (orange) and a 2.40 for references (green). The thesis with the highest thesis grade (last column on the right) was scored with a 3.00 for every single criterion. Figure 5 visualises that students can obtain nearly insufficient thesis grades even if their use of *references* is scored above 2.0 on average (i.e. green bars are visible on the far-left side of the figure). Additionally, Figures 4 and 5 both reveal that scores above 2.0 for *title, abstract*, *scope* and/or *spelling and grammar* are apparently not required to obtain excellent thesis grades (i.e. yellow bars are visible in the excellent bin of Figure 4 and on the far-right side of Figure 5). However, students never receive excellent thesis grades if their *professional attitude* and *structure* are scored with a 2.0 or below (i.e. only green bars are visible in the excellent bin of Figure 4 and on the far-right side of Figure 5).

Table 2. Regression analysis of thesis grade bins and criteria scores adjusted for interdependence between criteria scores. The regression analysis is performed separately for three bins; sufficient thesis grades (left column, <7.0; $n = 48$), good thesis grades (grades ≥ 7.0 to <8.5; $n = 158$) and excellent thesis grades (grades ≥ 8.5; $n = 55$).

		Sufficient thesis grade bin (≤7.0)					Good thesis grade bin (>7.0 to ≤8.0)					Excellent thesis grade bin (>8.0)				
	Rubric criterion	N[a]	S.D.	Mean	r	R^2_{adj}[b]	N[a]	S.D.	Mean	r	R^2_{adj}[b]	N[a]	S.D.	Mean	r	R^2_{adj}[b]
Opening	Title	93	0.29	2.00	0.10	0.00	144	0.47	2.28	0.21	0.04	59	0.48	2.64	0.17	0.01
	Abstract	93	0.38	1.89	0.14	0.01	142	0.52	2.36	0.29	0.08**	59	0.47	2.75	0.37	0.13*
	Introduction	93	0.43	2.03	0.25	0.05	144	0.46	2.53	0.39	0.15**	59	0.32	2.82	0.24	0.04
	Scope	85	0.54	1.89	0.33	0.10*	139	0.52	2.41	0.11	0.00	57	0.45	2.78	0.29	0.07*
Content	Structure	93	0.37	1.89	0.48	0.22**	144	0.38	2.53	0.48	0.23**	59	0.23	2.84	0.24	0.04
	Discussion	92	0.40	1.92	0.40	0.15**	141	0.39	2.36	0.54	0.29*	58	0.32	2.81	0.32	0.09
	Scientific quality	92	0.34	2.00	0.55	0.30**	144	0.32	2.55	0.50	0.25**	59	0.22	2.87	0.38	0.13*
Presentation	Spelling and grammar	91	0.50	2.16	0.24	0.05	140	0.49	2.58	0.27	0.07*	59	0.41	2.80	0.27	0.06
	Style	93	0.35	1.97	0.42	0.17**	144	0.36	2.48	0.39	0.15*	59	0.30	2.81	0.38	0.13*
	Length and layout	91	0.44	2.03	0.22	0.04	144	0.45	2.51	0.35	0.12**	58	0.34	2.79	0.20	0.02
	Figures	93	0.42	1.99	0.33	0.10*	144	0.50	2.41	0.38	0.14**	59	0.30	2.82	0.27	0.06
	References	93	0.37	2.22	0.27	0.06*	144	0.37	2.69	0.28	0.07**	59	0.22	2.88	0.16	0.00
Attitude	Professional attitude	89	0.39	2.08	0.53	0.27**	141	0.34	2.64	0.39	0.15*	57	0.18	2.91	0.10	0.00

Note. N is total number of cases, S.D. is standard deviation, R^2_{adj} is explained variance of thesis grade adjusted for the number of predictors

[a]The number of criteria scores differs per rubric category because of missing data.

[b]*$p < .01$; **$p < .001$.

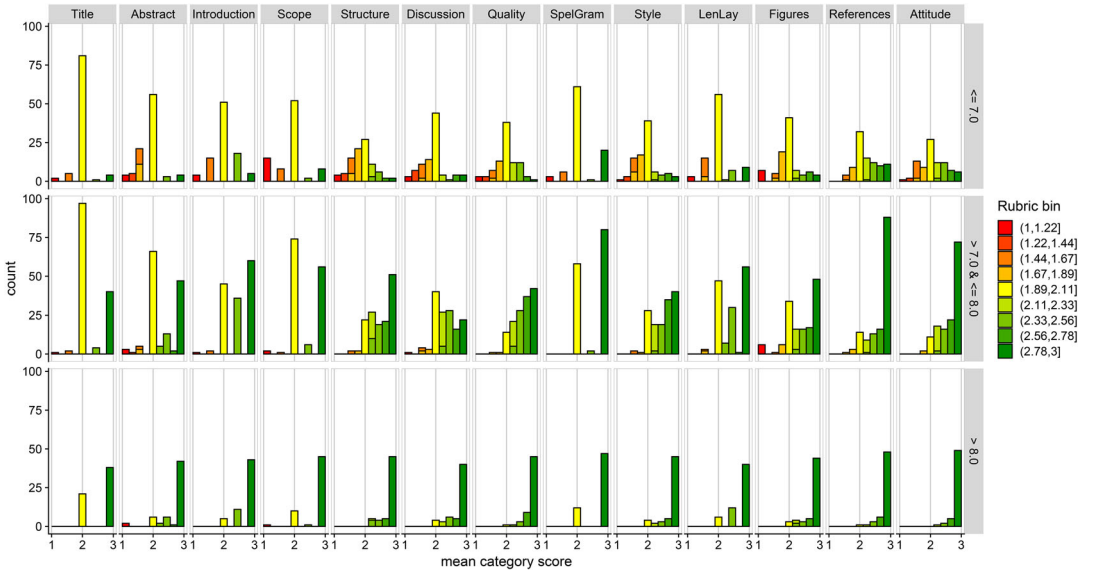

Figure 4. Frequency distribution of criteria scores of the 13 rubric categories for theses graded as sufficient (≤7.0), good (<7.0 and ≤8.0), and excellent (>8.0). Criteria scores are color coded from red (criteria scores from 1 to 1.22) to green (criteria scores from 2.78 to 3).

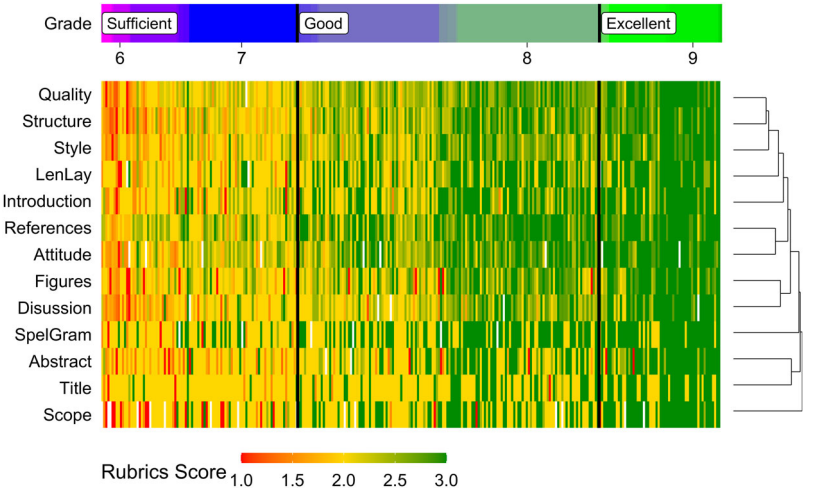

Figure 5. Average rubric criteria scores (rows) of all 318 assessed biology bachelor theses (columns). The 318 bachelor theses are sorted in columns from lowest thesis grade (left) to highest thesis grade (right) with thick black vertical lines marking the boundaries of the three separate grade bins: sufficient (≤7.0), good (<7.0 and ≤8.0), and excellent (>8.0). The criteria scores of each thesis are presented with color codes (red = sufficient, yellow = good and green = excellent). Criteria scores that were missing are marked in white. The criteria are ordered based on the similarity in scores, as indicated by the dendogram on the right. Criteria that are scored similarly, such as quality and structure or references and attitude, are directly connected in this dendogram. The dendogram is calculated based on Euclidean distance hierarchically clustering.

Discussion

Bachelor theses are essential for the completion of many study programmes and it is logically relevant that assigned grades reflect the quality of the students' work and attitude in the best possible manner. Our quantitative analysis provides insight in the assessment of bachelor theses with the aid of rubrics. We demonstrate the significance of each rubric criterion on the overall thesis grades, and separately on the grades of thesis deemed sufficient, good or excellent.

Theses that are assessed with the same grade can still vary in their average rubric scores and *vice* versa; identical rubric scores can result in different grades, depending on the relative contribution the examiners assign to different criteria. One explanation of the variance in total rubric scores is that some of the rubric categories appear more significant in the assessment than others. Indeed, we demonstrate that *scientific quality* and *structure* are better predictors for thesis grade than other criteria. The high predictability of *scientific quality* for assigned grade was expected since examiners are encouraged to put relatively more weight in their assessment to this criterion together with *professional attitude*. Unfortunately, we cannot assess to what extent this encouragement contributed to the high level of predictability of *scientific quality* for thesis grade. The result is nonetheless in agreement with a study by Lundgren, Halvarsson, and Robertsson (2008), who show that examiners mainly assess students' ability to compare results of different studies; one of the sub-criteria of *scientific quality* in the rubric of the current study. The other sub-criteria defined for *scientific quality* are 2) demonstration of a high level of knowledge, 3) adequate discussion of the results, 4) well explained and supported arguments, and 5) clear distinction of facts and hypotheses. It would not be surprising if these sub-criteria would also be the highest predictors for thesis grade if examiners were not informed of its importance. Most examiners in this study are tenured scientists and expected to assess theses from a researcher's perspective with focus on the students' line of reasoning. The experience of the examiner appears an important factor for assessments (Kapborg and Berterö 2002; van der Vleuten et al. 2012).

This study also reveals that some criteria are worse or better predictors for sufficient grades than for good or excellent grades, suggesting prioritization in the examiner's mind. Interestingly, *professional attitude* strongly predicts grades of sufficient theses. Bachelor theses also were never deemed excellent if a student scored low for *professional attitude,* suggesting that attitude is an important criterion for thesis assessment. The *professional attitude* includes sub-criteria on: 1) required feedback, 2) response to feedback, 3) enthusiasm and commitment, 4) fulfilling agreements and 5) fulfilling deadlines. Thus, we propose that if the content of the thesis is only just sufficient, but the student is independent, committed, enthusiastic, adjusts feedback and fulfils agreements, examiners are prone to lift the grade from sufficient to good. The sub-criteria of *professional attitude* on commitment, agreements and deadlines are independent of the quality of the thesis product itself and one can debate whether these should be considered when assessing theses. However, the first sub-criteria on required feedback discusses whether the thesis is a true product of the student and this level of 'self-regulatory ownership' is regarded essential for students in higher education (Prins, de Kleijn, and van Tartwijk 2017).

Remarkably, we found that the *abstract* is a better predictor for excellent grades than for sufficient or good grades. We hypothesise that, because the abstract is the first text the examiner reads, the quality of the abstract provides the first impression of the work that may set the tone for the rest of the assessment, consciously or unconsciously. Alternatively, as the abstract is often the finishing touch of a thesis, it is only considered by the examiner in the final product, if the rest of the thesis and prior drafts already appeared to be excellent.

Nonetheless, we note that all criteria scores are relatively less predictive for the assessment of excellent theses compared to theses with sufficient or good grades. This can be explained by the distribution of rubric scores, showing that most criteria of excellent theses are assessed with the highest possible score of 3.0 (see Figure 4). Thus, no further differentiation in rubric score can be made between criteria that only 'just fulfilled' the description of the highest rubric category or criteria that 'outreach' this description by far. Moreover, as the total number of cases is relatively low in the excellent grade bin, distinguishing rubric scores between 'just excellent grade' theses scored with an 8.0 and with 'highly excellent theses' scored with a 9.0 or 10.0 is also hampered by restrictions in statistical power. Alternatively, factors that are not included in the rubric play a role in the assessment.

Assessment procedure

The significance of the rubric criteria provides insight into the criteria that examiners deem relevant. However, this study does not disclose the implicit decisions that examiners make while assessing. We therefore encourage additional studies wherein examiners explain if and how they assign thesis grades with the aid of rubrics. This might reveal whether additional implicit factors (not described in the rubric) are considered during assessment and might explain if and how the rubric is used when deciding upon an overall grade. For example, examiners may fully rely on the rubric to determine the thesis grade, but it might also be that examiners simply use rubrics to justify the grade they already have in mind (Bloxham, Boyd, and Orr 2011; Timmerman et al. 2011). Most likely, our dataset covers the whole spectrum in between these extremes. It is naturally expected that examiners have different assessment approaches, and this might be particularly true for the biology theses of the current study that are assessed by examiners with diverse experiences in education, supervision and assessment, and are from diverse disciplines with each their own habits and cultures. Differences in final thesis grades have already been confirmed for external and internal examiners (Williams and Kemp 2019), for examiners with low and high affiliation with the student (de Kleijn et al. 2012), and for examiners who did or did not supervise the student themselves (Lundgren, Halvarsson, and Robertsson 2008). It appears that during assessment inexperienced examiners are more focused on each separate institutional criterion and that experienced examiners consider the interrelatedness of the criteria more often (Adnan and Bulut 2014; Kiley and Mullins 2004; Sadler 2009).

It would be interesting to study whether the rubric criteria scores are also different between these examiners and whether experienced examiners give more or less weight to certain criteria in comparison with inexperienced examiners. We should, however, stress that we do not recommend using a fixed formula to calculate a final thesis grade from criteria scores. We encourage to leave the overall judgment of theses to the supervisors since: i) they are the experts, ii) they can assess the overall quality and consider how criteria congregate and iii) they can judge how relevant each criterion is in their specific discipline. Nonetheless, in order to assess theses more similarly, we encourage moderation sessions wherein examiners explain to each other their understanding of the rubric criteria and discuss how they decide upon the final grade (Grainger, Adie, and Weir 2016).

Practical implications

The current study suggests that all criteria matter, showing that students should pay attention to every criterion. However, the regression analysis suggests that students should prioritise working on *scientific quality, structure* and *professional attitude* above criteria such as correct use of *spelling and grammar* and a catchy *title* that justifies its content. Thus, students are advised to be enthusiastic and committed (*professional attitude*) and explain, discuss and integrate the results (*scientific quality*) in a clear, consistent and structured manner (*structure*).

Supervisors are similarly advised to put emphasis on the most significant criteria and give extra feedback and guidelines to students on the *structure* and *scientific quality* of a thesis. Supervisors can improve the overall thesis quality of their students by giving extra guidance on criteria that are seldom scored as excellent such as *title*, *abstract* and *scope*.

Regression analyses are of specific interest for boards of examiners and education managers, as they can help in assessing whether the aims of their curriculum are achieved and whether the criteria they find of highest importance are most predictive for assigned grades. Fortunately, *scientific quality* is thought to be the most important category according to the rubric developers and also appears the best predictor for thesis grade. Similarly, *professional attitude* is highlighted as important and appears to be highly predictive for sufficient thesis grades as well. We do, however, not know if and how the highlights in the rubric assessed in this study have contributed to

the assessments. Either way, these results suggest that examiners primarily assess on the criteria that are also found to be most relevant by the board of examiners and institutional directives. We thus recommend performing similar analyses from time-to-time to estimate whether the assessment reflects the quality of the students' work and validate if assignments are mainly assessed on the criteria that are considered to be most relevant.

Conclusions

This explorative study presents a quantitative analysis from a large dataset of rubric scores and the corresponding thesis grades. The difficulty of assigning thesis grades is that many skills need to be considered to provide one single overall grade that fits all. We show that all rubric criteria scores are predictive for thesis grades, suggesting that all criteria are relevant for assessment. Furthermore, thesis grades are best distinguished by *scientific quality* and *structure*. In addition, the distinct analysis of criteria scores for separate grades showed that some criteria appear more predictive for low grades than for high grades. Information on these criteria might support students to prioritise on the most significant criteria, guide supervisors to give their students extra advice on these criteria and help education managers and accreditation committees to evaluate if the assessments are in agreement with the criteria they find of highest relevance.

Acknowledgements

We thank Dr. Margot Koster and Prof. Dr. Johannes Boonstra for their feedback and suggestions, and the Ethics Review Board of the Faculty of Sciences and Geosciences at Utrecht University for reviewing our proposal and advice. We also wish to acknowledge the members of the BSc Biology biannual theses quality examination committee, from whom assessments the idea for this study sparked and last but not least, all our thesis supervisors and examiners.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Marjolein Haagsman http://orcid.org/0000-0003-4000-1959
Basten Snoek http://orcid.org/0000-0001-5321-2996
Frans Prins http://orcid.org/0000-0002-7898-2978
Martijn van Zanten http://orcid.org/0000-0002-2810-7374

References

Adnan, K., and O. Bulut. 2014. "Crossed Random-Effect Modeling: Examining the Effects of Teacher Experience and Rubric Use in Performance Assessments." *Eurasian Journal of Educational Research* 14 (57): 1–28. doi:10.14689/ejer.2014.57.4.

Andrade, H., and Y. Du. 2005. "Student Perspectives on Rubric-Referenced Assessment." *Practical Assessment, Research and Evaluation* 10 (3): 1–11. doi: 10.7275/g367-ye94

Bloxham, S., P. Boyd, and S. Orr. 2011. "Mark my Words: The Role of Assessment Criteria in UK Higher Education Grading Practices." *Studies in Higher Education* 36 (6): 655–670. doi:10.1080/03075071003777716.

Bourke, S., and A. P. Holbrook. 2013. "Examining PhD and Research Masters Theses." *Assessment & Evaluation in Higher Education* 38 (4): 407–416. doi:10.1080/02602938.2011.638738.

de Kleijn, R. A. M., M. T. Mainhard, P. C. Meijer, A. Pilot, and M. Brekelmans. 2012. "Master's Thesis Supervision: Relations between Perceptions of the Supervisor-Student Relationship, Final Grade, Perceived Supervisor Contribution to Learning and Student Satisfaction." *Studies in Higher Education* 37 (8): 925–939. doi:10.1080/03075079.2011.556717.

de Vries, A., and B. D. Ripley. 2016. *gdendro: Create dendrograms and tree diagrams using "ggplot2"* (0.1-20). https://github.com/andrie/ggdendro

Goulden, N. R., and C. J. G. Griffin. 1995. "The Meaning of Grades Based on Faculty and Student Metaphors." *Communication Education* 44 (2): 110–125. doi:10.1080/03634529509379003.

Grainger, P., L. Adie, and K. Weir. 2016. "Quality Assurance of Assessment and Moderation Discourses Involving Sessional Staff." *Assessment & Evaluation in Higher Education* 41 (4): 548–559. doi:10.1080/02602938.2015.1030333.

Grömping, U. 2006. "Relative Importance for Linear Regression in R: The Package Relaimpo." *Journal of Statistical Software* 17 (1): 1–27. doi:10.18637/jss.v017.i01.

Kapborg, I., and C. Berterö. 2002. "Critiquing Bachelor Candidates' Theses: Are the Criteria Useful?" *International Nursing Review* 49 (2): 122–128. doi:10.1046/j.1466-7657.2002.00123.x.

Kiley, M., and G. Mullins. 2004. "Examining the Examiners: How Inexperienced Examiners Approach the Assessment of Research Theses." *International Journal of Educational Research* 41 (2): 121–135. doi:10.1016/j.ijer.2005.04.009.

Lundgren, S. M., M. Halvarsson, and B. Robertsson. 2008. "Quality Assessment and Comparison of Grading between Examiners and Supervisors of Bachelor Theses in Nursing Education." *Nurse Education Today* 28 (1): 24–32. doi:10.1016/j.nedt.2007.02.009.

Panadero, E., J. Alonso-Tapia, and E. Reche. 2013. "Rubrics vs. Self-Assessment Scripts Effect on Self-Regulation, Performance and Self-Efficacy in Pre-Service Teachers." *Studies in Educational Evaluation* 39 (3): 125–132. doi:10.1016/j.stueduc.2013.04.001.

Pedersen, T. L. 2019. *patchwork: The composer of plots* (R-package version 1.0.0). https://cran.r-project.org/package=patchwork

Prins, F. J., R. de Kleijn, and J. van Tartwijk. 2017. "Students' Use of a Rubric for Research Theses." *Assessment & Evaluation in Higher Education* 42 (1): 128–150. doi:10.1080/02602938.2015.1085954.

Reddy, Y. M., and H. Andrade. 2010. "A Review of Rubric Use in Higher Education." *Assessment & Evaluation in Higher Education* 35 (4): 435–448. doi:10.1080/02602930902862859.

Revelle, W. 2019. *psych: Procedures for psychological, psychometric, and personality Research* (R package version 1.9.12). https://cran.r-project.org/package=psych

Sadler, D. R. 2009. "Indeterminacy in the Use of Preset Criteria for Assessment and Grading." *Assessment & Evaluation in Higher Education* 34 (2): 159–179. doi:10.1080/02602930801956059.

Tai, J., R. Ajjawi, D. Boud, P. Dawson, and E. Panadero. 2018. "Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work." *Higher Education* 76 (3): 467–481. doi:10.1007/s10734-017-0220-3.

Timmerman, B. E., D. C. Strickland, R. L. Johnson, and J. R. Payne. 2011. "Development of a 'Universal' Rubric for Assessing Undergraduates' Scientific Reasoning Skills Using Scientific Writing." *Assessment & Evaluation in Higher Education* 36 (5): 509–547. doi:10.1080/02602930903540991.

van der Vleuten, C. P. M., L. W. T. Schuwirth, E. W. Driessen, J. Dijkstra, D. Tigelaar, L. K. J. Baartman, and J. van Tartwijk. 2012. "A Model for Programmatic Assessment Fit for Purpose." *Medical Teacher* 34 (3): 205–214. doi:10.3109/0142159X.2012.652239.

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, et al. 2019. *gplots: Various R programming tools for plotting data* (3.0.3). https://github.com/talgalili/gplots

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag. https://doi.org/978-3-319-24277-4.

Wilke, C. O. 2019. *cowplot: Streamlined plot theme and plot annotations for ggplot2*. https://github.com/wilkelab/cowplot

Williams, L., and S. Kemp. 2019. "Independent Markers of Master's Theses Show Low Levels of Agreement." *Assessment & Evaluation in Higher Education* 44 (5): 764–771. doi:10.1080/02602938.2018.1535052.

Appendices

Appendix 1. Rubric used for assessing Biology bachelor theses at Utrecht University.

Criteria	Insufficient	Sufficient	Good
Title	• Title does not justify the content • Title is too long	• Title justifies the content	• Title justifies the content and is catching
Abstract	• Abstract is missing • Abstract is unclear • Abstract lacks, or gives too much, irrelevant information	• The abstract is present • In general the abstract is clear • The abstract contains all important elements	• Abstract is present • Abstract contains a clear, concise and balanced description of all important elements.
Introduction and scope of thesis	• The introduction contains incomplete information • The introduction contains a lot of irrelevant information • The scope of the thesis is missing • The scope of the thesis is unclear	• The introduction contains sufficient information • The introduction is functional • Most of the presented information is relevant • Some repetition is present, but this is not disturbing • The scope of the thesis is clear	• The introduction contains sufficient relevant, information to put the scope in a broader perspective • The introduction is catchy and inviting to continue reading • The scope of the thesis is well defined
Structure	• The order of paragraphs is illogical. • Transitions between the paragraphs are lacking • The different sections are not in balance • Not all of the information is consistent with the scope • Information is frequently repeated • The line of thought is difficult to follow	• The structure and order of most paragraphs is sound • Transitions between paragraphs are mostly logical • Most sections are well balanced • Most of the information is consistent with the scope • Information is sometimes unnecessarily repeated • The line of thought is clear	• The structure and order of the paragraphs is sound • Smooth transitions between paragraphs • The sections are well balanced • The information is consistent with the scope • Only if necessary, the information is repeated • The line of thought is clear and well structured
Discussion/ Conclusion	• The discussion is missing or too short • The discussion lacks depth • The discussion does not go back to the scope of the thesis • No conclusions are drawn • The conclusions are not supported by the presented findings	• The discussion ties together loose ends • The discussion has limited depth • The discussion goes back to the scope of the thesis • Conclusions are drawn in line with the presented findings	• Discussion demonstrates a good overview of the subject of the thesis • In depth discussion • The discussion goes back to the scope of the thesis • Conclusions are drawn in line with presented findings
Scientific quality	• The thesis does not demonstrate sufficient knowledge of the field • Results from the referred articles are not discussed correctly • Arguments are not supported by evidence • It is difficult to distinguish facts from hypotheses • Information is not well integrated: the text often reads as loose pieces of information	• The thesis shows sufficient knowledge of the field in most parts • Most important results from the references are discussed correctly • Some arguments are supported by evidence • Facts and hypotheses are not always easy to distinguish from each other • Limited integration of information	• The thesis shows a high level of knowledge and understanding of the field • Results from the references are adequately discussed • Arguments are well explained and scientifically supported • Facts and hypotheses can easily be distinguished • Results from different papers are integrated well into new insights
Spelling and grammar	• Disturbing spelling or grammar mistakes, which complicates understanding of the text	• There are errors in spelling and/or grammar, but this does not prevent understanding the text.	• None, or very few errors in grammar or spelling
Style	• Writing style is hardly scientific	• In general the writing style is scientific	• Style is concise, and scientific. It is a pleasure to read the text

(*continued*)

Appendix 1. Continued.

Criteria	Insufficient	Sufficient	Good
	• Many sentences are not structured properly, or too long or too short • Little variation in vocabulary • Important terms are not explained • Bad use of punctuation	• Few sentences are badly structured, but this is not too disturbing • In general a sufficient vocabulary is used • Important terms are explained • Use of punctuation is usually correct and makes the text easier to follow	• Sentences flow smoothly and are not too long or too short • Rich vocabulary • Important terms are well explained • Good use of punctuation
Length/Layout	• The length is too short / too long • The layout is sloppy and not appealing	• The length is as required • The layout is uniform and organized	• The length is as required • The layout is uniform and well organized
Figures	• Insufficient number of figures • Figures are often of bad quality • Too many irrelevant figures • Figure legends are missing or badly formulated • The figures are not referred to in the text or the reference is wrong • A clear explanation of the figure in the text is missing	• Sufficient number of figures • Most figures are of good quality • Most figures are relevant • Figure legends are adequate to understand the figure • Good reference to the figures in the text • Most figures are explained well in the text	• Sufficient number of figures, that support the text • Figures are of good quality • Figures are relevant and when needed adjusted or specifically designed for the thesis • Figure legends are complete • Good reference to the figures in the text • Figures are explained well in the text
References	• Insufficient number of references (<20) • An insufficient number of primary literature is used • Finding and selecting references was mainly done by the supervisor • None, or incorrect referral to references in the text • The reference list is incomplete or sloppy	• Sufficient number of references (>20) • A sufficient amount of primary literature was used • Most sources were found by the student • Most of the articles are correctly referred to in the text • The reference list is largely complete, clear and uniform, with some minor mistakes	• Sufficient number of references (>20) • A sufficient amount of primary literature was used • All references were found by the student • The articles are correctly referred in the text • The reference list is complete, clear and uniform
Professional attitude	• A lot of supervision was required • Minimal improvement after feedback • The student did not put enough effort into the thesis work • The student failed to follow the agreements • The student failed to meet the deadlines	• Regular feedback sessions were required • Feedback led to reasonable improvements • The student was committed • The student fulfilled the agreements • The students met the agreed deadline	• Minimal feedback was required, the thesis is a true product of the student • Response to feedback yielded excellent improvements • The student was enthusiastic and committed • The student fulfilled the agreements • The student met the agreed deadline