Received 30 August 2022
Revised 7 January 2023
31 March 2023
Accepted 4 May 2023

Ensuring bachelor's thesis assessment quality: a case study at one Dutch research university

Ya-Ping (Amy) Hsiao
Tilburg University, Tilburg, The Netherlands
Gerard van de Watering
Eindhoven University of Technology, Eindhoven, The Netherlands
Marthe Heitbrink
University of Amsterdam, Amsterdam, The Netherlands
Helma Vlas
University of Twente, Enschede, The Netherlands, and
Mei-Shiu Chiu
National Chengchi University, Taipei City, Taiwan

Abstract

Purpose – In the Netherlands, thesis assessment quality is a growing concern for the national accreditation organization due to increasing student numbers and supervisor workload. However, the accreditation framework lacks guidance on how to meet quality standards. This study aims to address these issues by sharing our experience, identifying problems and proposing guidelines for quality assurance for a thesis assessment system.

Design/methodology/approach – This study has two parts. The first part is a narrative literature review conducted to derive guidelines for thesis assessment based on observations made at four Dutch universities. The second part is a case study conducted in one bachelor's psychology-related program, where the assessment practitioners and the vice program director analyzed the assessment documents based on the guidelines developed from the literature review.

Findings – The findings of this study include a list of guidelines based on the four standards. The case study results showed that the program meets most of the guidelines, as it has a comprehensive set of thesis learning outcomes, peer coaching for novice supervisors, clear and complete assessment information and procedures for both examiners and students, and a concise assessment form.

Originality/value – This study is original in that it demonstrates how to holistically ensure the quality of thesis assessments by considering the context of the program and paying more attention to validity (e.g. program curriculum and assessment design), transparency (e.g. integrating assessment into the supervision process) and the assessment expertise of teaching staff.

Keywords Quality assurance, Accreditation, Thesis assessment

Paper type Research paper

The authors are grateful to the reviewers for their thorough review and valuable feedback, which allowed the authors to improve the quality of the manuscript. The authors appreciate the time and effort they put into the review process.

Funding: This work was supported by National Chengchi University (DZ15-B4). The funder only provides financial support and does not substantially influence the entire research process, from study design to submission. The authors are fully responsible for the content of the paper.

Higher Education Evaluation and Development
Vol. 18 No. 1, 2024
pp. 2-16
Emerald Publishing Limited
2514-5789
DOI 10.1108/HEED-08-2022-0033

Introduction

According to data from the universities of the Netherlands, the number of bachelor's students at Dutch research universities has been steadily increasing from 2015 to 2021 [1], leading to increased workload for teaching staff due to the need for greater supervision of students [2]. This increased supervision is particularly evident in the supervision of students' final projects. In the Netherlands, students can begin working on their final projects in the final year of their program's curriculum once they pass the first-year diploma (the so-called Propaedeutic phase based on a positive binding study advice, BSA), earn a required number of European Credit Transfer and Accumulation System (ECTS) credits and meet other requirements. A bachelor's degree is awarded when a student has "demonstrated by the results of tests, the final projects, and the performance of graduates in actual practice or in postgraduate programmes" (The Accreditation Organisation of the Netherlands and Flanders [Nederlands-Vlaamse Accreditatieorganisatie], hereinafter abbreviated as the NVAO, 2018, p. 34).

At Dutch research universities, the most commonly used final project is a thesis. Among various definitions, the one given by The University of Twente best describes the shared genre of a bachelor's thesis used to assess the achievement at the exit level of a study program at Dutch research universities (University of Twente, 2019):

> The Bachelor's thesis is the culmination of the Bachelor's programme. A Bachelor's thesis is carried out in the form of a research project within a department. It is an opportunity to put the knowledge learned during the programme into practice. The Bachelor's thesis is used to assess the student's initiative and their ability to plan, report and present a project. The difficulty level of the thesis is described by the attainment targets of the programme and the modules followed up until that moment. Students work independently on a Bachelor's thesis or Individual Assignment (IOO) under the guidance of a supervisor.

This definition highlights the pedagogical value of the thesis (i.e. the opportunity to carry out an independent project) and the purpose of thesis assessment (i.e. to determine the extent to which the intended learning outcomes have been achieved). While this definition acknowledges the importance of a bachelor's thesis, relatively little research has been done on examining the quality of undergraduate thesis assessment (Hand and Clewes, 2000; Shay, 2005; Webster *et al.*, 2000; Todd *et al.*, 2004), let alone in the Dutch context where thesis supervisors and examiners of bachelor's students are experiencing an increasing workload.

In recent years, the Dutch government has placed increasing emphasis on assessment quality in higher education (Inspectorate of Education [Inspectie van het Onderwijs], 2016). The NVAO has established the Assessment Framework for the Higher Education Accreditation System of the Netherlands (hereinafter abbreviated as the Framework, NVAO, 2018). The standards for the accreditation of initial and existing study programs emphasize whether a program has established an adequate student assessment *system* that appropriately assesses the intended learning outcomes (NVAO, 2018). According to the quality standards of the Framework, thesis assessment should be valid, reliable, transparent and independent. Assessment literature in the higher education context has defined these criteria as follows (e.g. Biggs and Tang, 2007; Bloxham and Boyd, 2007). Validity refers to the extent to which an assessment accurately measures what it is intended to measure. Reliability refers to the consistency of the assessment results, or how well they accurately reflect a student's actual achievement level. Transparency is the clarity and specificity with which assessment information is communicated to both students and examiners. Independency is a necessary condition for ensuring the validity and reliability of an assessment, as it requires that examiners remain objective in the assessment process.

Despite the inclusion of these standards in the Framework (NVAO, 2008), official guidance on establishing a quality system of assessing graduation projects that test achievement of the

exit level of a study program at Dutch research universities is limited. As assessment practitioners (the first four authors of this article), we have found that it is often unclear for a program's curriculum and/or management team to establish appropriate thesis assessment procedures at the undergraduate level that meet the NVAO's quality standards. We hope that our experience can provide valuable insights and guidance for programs seeking to ensure quality assurance for thesis assessment.

Aims and research questions

The purpose of this study is to share our experience and the challenges we faced during internal and external quality assurance processes of thesis assessment. Based on these challenges, we conducted a narrative literature review to develop a set of guidelines for ensuring thesis assessment quality that aligns with the four standards outlined in the Framework (NVAO, 2018): (1) intended learning outcomes, (2) teaching and learning environment, (3) assessment and (4) achieved learning outcomes. To illustrate the application of these guidelines, we present a case study of bachelor's thesis assessment practices at one Dutch research university.

The research questions we aim to answer in this study are as follows:

RQ1. What are the guidelines for ensuring the quality of thesis assessment procedures that meet the standards specified in the Framework?

RQ2. How can these guidelines be applied to evaluate the quality of thesis assessment in a study program?

It is important to note that this study is limited to the context of four Dutch research universities, where we encountered common issues during internal quality assurance processes of thesis assessment. Our goal is to share our experience and offer insights that could be useful to other institutions seeking to ensure the quality of thesis assessment. We do not intend to assume that these problems are present at all Dutch research universities.

Problems and guidelines in meeting the four standards

According to the didactic principle of constructive alignment (Biggs and Tang, 2007), which is commonly used in Dutch higher education, the three education processes, teaching, learning and assessment, should be aligned with the intended learning outcomes. We begin with Standards 1 and 2, which set out the conditions under which thesis assessment takes place, and then we place more emphasis on Standards 3 and 4, which focus on the quality criteria for thesis assessment.

Standard 1: intended learning outcomes

Problems. To ensure that a study program meets Standard 1 of the Dutch Qualification Framework (NLQF, 2008), the intended learning outcomes for graduates in specific subject areas and qualifications are typically developed using the Dublin Descriptors (Bologna Working Group, 2005), which provide generic statements of competencies and attributes. However, it is often assumed that a thesis should assess all of these program learning outcomes (PLOs) since it is intended to evaluate the achieved learning outcomes at the exit level. Unfortunately, these PLOs can be global and unclear, which can confuse and hinder students from trying to understand the expectations for thesis assessment. Our observation is that programs often utilize PLOs as thesis learning outcomes (TLOs), although a thesis is not equivalent to the entire program curriculum.

Guidelines. According to Biggs and Tang (2007), it is important for teachers to first clearly define the learning outcomes before designing instructional activities to guide students toward achieving them. In addition, the outcomes at the program and course levels (i.e. a

thesis is also a course) should also be constructively aligned, and the course-level outcomes should be specific to the context of the course. Therefore, to design effective thesis activities (such as supervision) and develop assessment criteria, it would be more pedagogically valuable to formulate *thesis-specific* learning outcomes and explain how they contribute to the PLOs and Dublin Descriptors, rather than directly using the PLOs for thesis assessment.

In addition, a thesis course often involves most of the teaching staff in the program. Therefore, it is important to establish clear and *specific* expectations for what students should achieve at the end of a bachelor's thesis course (Willison and O'Regan, 2006; Todd *et al.*, 2004), such as the scope and type of research (e.g. scaffolded or self-initiated), integrating disciplinary knowledge and research skills from earlier program curriculum, demonstrating critical thinking through well-supported arguments and developing independent learning skills for future work (Willison and O'Regan, 2006).

Standard 2: teaching-learning environment

Problems. According to Standard 2 of the Dutch Qualification Framework (NLQF, 2008), the quality of the teaching and learning environment should be designed to help students achieve the intended learning outcomes of the program curriculum. However, our experience has revealed problems in this area. In informal discussions with thesis supervisors, we have found that students often report a lack of preparedness for a bachelor's thesis, as they have not been adequately taught or practiced certain academic and research skills such as communication, information seeking and methodologies. Conversely, many teachers in the program believe they have covered these skills in their courses. Furthermore, during thesis calibration sessions, we have observed that novice examiners lack expertise due to insufficient experience in research education, a lack of training as thesis examiners, and unclear instructions on thesis assessment procedures.

Guidelines. To meet Standard 2, we recommend the following two guidelines. First, as suggested by research on curriculum alignment (Wijngaards-de Meij and Merx, 2018) and research skills development (Willison, 2012; Reguant *et al.*, 2018), the program-level curriculum design should arrange domain-specific subjects in a logical order and gradually develop students' research, communication and independent learning skills so that they are well prepared to work on the thesis. At the same time, universities should focus on converting teaching staff's research experience into research education expertise (Maxwell and Smyth, 2011) for the long term.

Second, the program should ensure the quality of the teaching staff because examiners' practices are crucial for the quality of thesis assessment (Golding *et al.*, 2014; Kiley and Mullins, 2004; Mullins and Kiley, 2002). According to the literature, thesis examiners should receive sufficient instructions and training on how to grade a thesis (Hand and Clewes, 2000; Kiley and Mullins, 2004). In addition, the university should provide teaching staff with written instructions to regulate and communicate thesis assessment procedures for supervisors, examiners and students, as well as assessment training on using the assessment forms and holding calibration sessions to achieve consistency in interpreting criteria and grade points. The literature on how supporting teaching staff in assessment practices contributes to consistency is discussed further in the section on Reliability.

Standards 3 and 4: student assessment and achieved learning outcomes

Validity. Ensuring validity starts with clearly defining what the assessment is intended to measure. According to the definition of validity and principle of constructive alignment (Biggs and Tang, 2007), thesis assessment should be aligned with learning outcomes.

Problems. We have identified two problems in this regard. The first problem is the use of a *generic* assessment form with a set of uniform criteria across different programs within the same department or school. We believe this practice does not follow the principle of constructive alignment (Biggs and Tang, 2007). In particular, the same assessment form cannot be used directly for different degrees (i.e. Bachelor, Master and PhD) based on the Dublin Descriptors. It would be difficult for a generic assessment form to assess the different levels of cognitive demand and skills required at each degree level. For example, the concept of "originality" is defined very differently at each degree level and this should be reflected in the assessment criteria.

The second problem is the quality of the assessment form itself. We have observed the following issues: (1) some criteria are not always directly relevant to the TLOs, (2) the assessment form only lists the names of criteria without defining them or providing specific indicators for each criterion, (3) it is unclear whether different criteria are given equal weight and (4) it is unclear how the final grade is determined (e.g. whether each criterion must be "sufficient" or "passing").

Guidelines. To address these problems, we recommend the following guidelines. The assessment criteria listed in the form should align with the TLOs and should describe the characteristics of student work that provide relevant, representative and important evidence of their attainment of the learning outcomes (Brookhart, 2013, 2018; Walvoord and Anderson, 2011). In addition to aligning the criteria with the outcomes, the quality of the criteria also affects what is actually being assessed. The criteria should avoid vagueness that leads to multiple interpretations of quality indicators (Biggs and Tang, 2007; Bloxham *et al.*, 2011; Hand and Clewes, 2000; Webster *et al.*, 2000). To ensure that the assessment measures what it is intended to measure, the criteria should meet the following five criteria (Brookhart, 2013, 2018; Walvoord and Anderson, 2011): they should be definable, observable, distinct from one another, complete and able to support descriptions along a continuum of quality.

Another important aspect of validity is the weighting of multiple assessment criteria. The weighting should reflect the relative importance of the criteria based on the disciplinary focus of the study program. For example, the criterion of "method and data analysis" might carry more weight in psychology than it would in philosophy.

Reliability and independency. Reliability is a necessary condition for validity and refers to the consistency of assessment results. Reliability is important because it allows us to confidently interpret and determine students' true performance on a thesis.

Independency between examiners is necessary to ensure the reliability (or objectivity) of the assessment process, as it helps prevent influence on each other's judgment. Independent grading is often specified in the Education and Examination Regulations of an institution.

Problems. Intra-rater reliability refers to the consistency of a single examiner's grading process over time. Inconsistencies may occur due to internal influences rather than true differences in student performance. We have observed inconsistencies in completed assessment forms, including discrepancies between comments and scores given by the same examiner across different student theses.

Inter-rater reliability, on the other hand, refers to the consistency of grading behavior between examiners. A lack of standard assessment procedures can lead to inconsistency in grading, such as a tendency for supervisors to assign higher grades than second examiners. In addition, we have observed the following three different assessment procedures used by examiners within a program (Hsiao and Verhagen, 2018):

(1) Analytical: Examiners assign a rating to each criterion and then determine a thesis grade based on the grading guidelines.

(2) Analytical and then holistic: Examiners assign a rating to each criterion and then determine a thesis grade based on the grading guidelines. If the thesis grade does not match the holistic judgment, examiners adjust the ratings of the criteria.

(3) Holistic and then analytical: Examiners hold an initial grade (in their mind) based on holistic judgment. Next, examiners assign a rating to each criterion and determine a thesis grade based on the grading guidelines. If the thesis grade is different from the initial grade, examiners adjust the ratings of the criteria to make sure that these two grades are the same.

Guidelines. To ensure intra-rater reliability, it is essential to clearly define each criterion to prevent multiple interpretations by examiners. Additionally, examiners should be provided with bias-reduction training (Wylie and Szpara, 2004) to make them aware of potential biases, such as supervisor bias (Bettany-Saltikov *et al.*, 2009; McQuade *et al.*, 2020; Nyamapfene, 2012), and to take actions to prevent them. During the grading process, examiners should also consistently revisit the established criteria and level descriptors to maintain consistency.

To improve inter-rater reliability, the literature suggests establishing standard assessment procedures and improving examiners' assessment practices (Hand and Clewes, 2000; Kiley and Mullins, 2004; Pathirage *et al.*, 2007). Standard assessment procedures should clearly outline the process for considering the relative importance of multiple criteria and the relative importance of various indicators within a criterion (Hand and Clewes, 2000; Bloxham *et al.*, 2016a; Pathirage *et al.*, 2007; Webster *et al.*, 2000). To improve examiners' assessment practices, common approaches include providing examiners with the following three processes (Sadler, 2013):

Prior to grading, to ensure consistent grading, examiners should have a shared understanding of the expectations for each criterion and score level. This can be achieved through the use of anchor or exemplar theses, which are previously graded theses that illustrate the characteristics of each score level (Osborn Popp *et al.*, 2009). Examiners can refer to these anchor theses as they grade to ensure that they are accurately distinguishing between the different score levels. It should also be clear to examiners how to complete the grading form and whether they are allowed to discuss with other examiners during the grading process (Pathirage *et al.*, 2007; Dierick *et al.*, 2002).

During the grading process, moderation refers to the process of two examiners arriving at a collective thesis grade (Bloxham *et al.*, 2016b). It is important to have clear instructions on how to control evaluative judgments and stay within reasonable limits during the moderation process. Examiners should also be informed of score resolution methods in case of large discrepancies between their scores, as averaging the scores may not be sufficient in such cases (Johnson *et al.*, 2005; Sadler, 2013). If a third examiner is involved in the moderation process, it should be clear who is qualified for this task and how their results are used to determine the final thesis grade (Johnson *et al.*, 2005).

As a "post-judgment" process, calibration is the act of ensuring that examiners grade student work against the agreed quality criteria and "how a particular level of quality should be represented" (Sadler, 2013, p. 6). It can be helpful to think of calibration as similar to checking the accuracy of a weighing scale by comparing it to a standard and making adjustments to bring it into alignment. In a similar vein, the thesis assessment form (including criteria and score-level descriptors) and examiners' assessment practices should be calibrated, particularly when there are significant changes in thesis assessment procedures. As noted by Sadler (2013), high-quality evaluative judgments also require the development of "calibrated" academics who serve not only as custodians of quality criteria and level standards but also as consultants for novice and short-term examiners. Calibration

can be implemented alongside the normal grading period as part of an internal quality assurance system (Andriessen and Manders, 2013; Bergwerff and Klaren, 2016).

Transparency. Transparency in assessment has received increasing attention in higher education in recent years (Bamber, 2015; Bell *et al.*, 2013; O'Donovan *et al.*, 2004; Price, 2005). It refers to making the perceptions and expectations of assessors, including requirements, standards and assessment criteria, known and understood by all participants, particularly students (O'Donovan *et al.*, 2004).

Problems. To ensure transparency in thesis assessment, it's not enough to only provide students with assessment forms and instructions on assessment procedures. Our observations indicate that without discussing the deeper meaning of criteria and standards, there is a risk of different interpretations by examiners and students.

Guidelines. To address this issue, it is important to foster shared understanding and promote assessment for learning and feedback on progress. This can be achieved by helping students develop their understanding of the quality criteria and standards through observation, discussion and imitation of good-quality theses (Malcolm, 2020). Using anchor theses (Orsmond *et al.*, 2002; Sadler, 1987) and involving students in peer review and grading of each other's theses using the criteria (O'Donovan *et al.*, 2004; Rust *et al.*, 2003) can be effective ways to do this.

To ensure transparency, supervisors should use the assessment form not only for thesis examination but also during supervising activities, and should clearly explain the criteria and score levels to their students using anchor theses for illustration (O'Donovan *et al.*, 2004; Rust *et al.*, 2003).

The guidelines for the four standards are summarized in Box 1 below.

Box 1. Overview of guidelines

Standard 1 – intended learning outcomes
- Formulate program-specific TLOs.

Standard 2 – teaching-learning environment
- Thesis assessment should be appropriate for the program curriculum and assessment plan.
- The program should ensure examiners' assessment expertise by providing training or instructions.

Standards 3 and 4 – student assessment and achieved learning outcomes

Validity
- TLOs, thesis supervision and thesis assessment should be constructively aligned.
- The assessment criteria should be clearly defined and meet quality requirements. The weighting of multiple criteria should reflect the relative importance of TLOs.

Reliability
- Intra-rater reliability: Examiners should revisit the established criteria to ensure consistency and strive to prevent any possible assessor bias.
- Inter-rater reliability: The program should establish assessment procedures and improve examiners' assessment practices.
  - The program should make assessment procedures consistent across examiners.
  - The program should improve examiners' assessment practices through the use of anchor or exemplary theses, moderation prior to and during assessment practices, and calibration after thesis assessment.

Transparency
- The program should inform students of what is expected of them and how their thesis will be assessed.
- The program should instruct supervisors to explicitly use the criteria during supervising activities.

Case study

To illustrate the application of these guidelines, we present a case study of a psychology-related bachelor's program at a Dutch research university. We chose to focus on this program because all of the authors have experience in quality assurance at various psychology programs. The documents for this case study were provided by one of the co-authors, who played a significant role in the quality assurance of assessment at the program. These documents include the program's learning outcomes, a thesis handbook, a thesis assessment form, grading instructions for examiners and a self-assessment report (which includes reflections on the four standards of the Framework and is required to be submitted to the NVAO before a site visit).

Four of the authors and the vice program director (as a self-reflection exercise) examined these documents and answered open-ended questions derived from the guidelines in Box 1. The findings were then structured based on the guidelines in Box 1.

Motivation for participating in this study

After receiving feedback from the previous NVAO site visit, the study program is currently working on improving two aspects of its assessment practices:

(1) Improving the quality of the assessment criteria to prevent multiple interpretations by examiners.

(2) Clearly defining the roles, tasks and responsibilities of supervisors (as the first examiner) and the second examiner.

The vice program director indicated that the assessment form is still in development and that it is a dynamic improvement process, based on examiners' accumulated experience and feedback from supervisors, examiners, students and assessment specialists.

Brief course descriptions of the Bachelor's thesis

In this thesis course, students perform a study that covers the entire empirical research cycle, from developing a specific research question to using theory to answer the question and testing the theory through data collection. They integrate knowledge from various disciplines and practice conducting research on a technology-related problem. Students may collaborate in groups for literature search or data collection, but they must formulate a specific question to be answered in their individually written bachelor's thesis.

Standard 1 – intended learning outcomes

The program committee has specified PLOs to state graduates' knowledge and skills in terms of seven competences:

(1) PLO1 – Competent in scientific disciplines

(2) PLO2 – Competent in doing research

(3) PLO3 – Competent in designing

(4) PLO4 – Use of a scientific approach

(5) PLO5 – Basic intellectual skills

(6) PLO6 – Competent in cooperating and communicating

(7) PLO7 – Take into account the temporal, technological and social context.

The thesis coordinator has formulated TLOs and related each outcome to multiple PLOs (indicated in brackets). Students are capable of performing the following activities under supervision:

(1) TLO1 – formulate a research question fitted to the problem and relevant scholarly literature (PLO1,2)

(2) TLO2 – conduct a literature search (PLO1,2,3,4,6)

(3) TLO3 – apply and modify relevant scientific theory in order to solve a technology-related problem (PLO1,2,4,5,7)

(4) TLO4 – make an adequate research design for empirical research (PLO2,3,4)

(5) TLO5 – apply relevant scientific methods for empirical research (PLO1,2,3,4,5)

(6) TLO6 – relate interpretation of data to theory and to design and/or policy recommendations (PLO1,2,3,4,5,7)

(7) TLO7 – individually write a scientific report (PLO5,6)

(8) TLO8 – reflect and think systematically (PLO5,6,7)

We conclude that TLOs contribute to the development of all seven competences outlined in the PLOs, as well as the five components of the Dublin Descriptors.

Standard 2 – teaching-learning environment

The bachelor's thesis builds upon the knowledge and skills developed in previous courses. According to the curriculum and program assessment plan, student skills progress from year 1 to 3 and are assessed through various types of assessment, such as presentations, reports and reflective writing. However, there is no specific learning trajectory for academic and research skills available.

To ensure student readiness for working independently on their thesis, students must have passed the propaedeutic phase and obtained a required number of ECTS upon enrolment in the bachelor's thesis course. They must also have passed the two methods courses.

Written instructions, including a detailed explanation of assessment procedures, criteria and rubrics, are provided in a thesis handbook for supervisors, examiners and students.

The program requires novice examiners to go through an "examiner internship" with senior examiners (mentors). They are guided and monitored by their mentors when assessing graduation theses in their first year of practice. They can directly approach mentors when encountering problems during supervision and assessment.

Standards 3 and 4 – student assessment and achieved learning outcomes

Validity. The thesis assessment form contains six criteria. All of the TLOs are assessed by multiple criteria. The assessed TLOs are indicated in brackets below.

(1) C1 – Abstract (TLO7,8)

(2) C2 – Introduction/Theory (TLO1,2,3,8)

(3) C3 – Method and results (TLO2,4,5,6)

(4) C4 – Discussion (TLO1,2,3,6,8)

(5) C5 – Writing style (TLO7)

(6) C6 – Process/Work attitude (TLO7,8)

Each criterion on the assessment form includes a short definition and a number of indicators, which are graded using a five-point rating scale (Poor–Insufficient–Sufficient–Good–Very good). It is required that qualitative comments be added to all of the criteria.

However, there are three issues with the assessment form:

(1) It is not clear how each criterion is weighted.

(2) It is not clear how the ratings of multiple indicators and criteria are aggregated to determine the total grade.

(3) Although a rating scale is provided, score-level descriptors are not available. It is not clear whether the indicators describe the "Very good" or "Sufficient" score level.

These issues correspond to areas that the program is currently working to improve, as mentioned at the beginning of this section.

Reliability. The program has established the following assessment procedures to ensure intra- and inter-rater reliability as well as to clearly define the role tasks and responsibilities of the examiners (see Motivation of participating in this study).

(1) New examiners receive a one-day training, in which they practice assessing theses based on the rubric, and discuss their practice results with senior examiners. They also receive guidance on how to use the criteria during the supervision process.

(2) The first and second examiners assess the thesis independently by using the same rubric and register their initial grading results *separately* to the administration system.

(3) It is obligatory for both examiners to hold a moderation meeting in order to arrive at collective grading results. In this meeting, they go through each criterion and discuss the differences. Then they register the collective results in the administration system, which generates the thesis grade.

(4) When the discrepancies between two examiners cannot be moderated during the meeting, both examiners register these in the administration system. Next, a subcommittee from the Examination Board is informed, which carries out additional grading. The members of the subcommittee are senior examiners who are often mentors assigned to the novice examiners during the examiner internship.

(5) There are no institution-wide guidelines on the moderation and calibration process. These quality assurance processes are organized by study programs. How they are implemented depends on the available resources, assessment expertise and time per study program.

(6) Although no calibration procedure is established, the subcommittee regularly regrades a sample of the borderline theses around the fail/pass grade, the theses with a resit, and theses for which the two examiners differ substantially in their initial grading. In addition, this subcommittee holds a regular plenary meeting to discuss their assessment practices and report their findings regularly to the Examination Board.

(7) After the assessment, both examiners and students are asked to fill out a survey to evaluate the use of rubric and the assessment procedures. The results are used for improving the quality of rubric.

These procedures are in line with most of our guidelines. Still, we suggest that the subcommittee systematically analyses their findings of regrading practices and acts on the

improvements in order to complete the quality assurance cycle. In addition, as lessons learned from one university, we highly recommend the Examination Board or the program to carry out a regular review of the completed assessment forms to detect whether there is any assessor bias in order to safeguard intra-rater reliability.

Transparency. The program has established clear guidelines on how to ensure transparency. At the beginning of the final project, an information session is organized to explain the supervision and assessment procedures and rules to students. It is made clear what the role tasks and responsibilities of supervisor, examiner and student are, in what way the thesis is assessed, and what is assessed (i.e. the criteria in the rubric). The criteria and indicators per criterion are explained in detail in this information session.

The program also makes it clear that the criteria should be used from the beginning and during the supervision activities, as well as in the assessment process. Supervisors are instructed to formulate feedback based on the criteria.

To sum up, this case study shows that their thesis assessment practices apply most of the guidelines suggested in this study.

Conclusion and discussion

This study presents problems encountered from a practitioner's perspective and derives guidelines from the literature to address these issues. These guidelines cover the entire education process, taking the context of the program into account. They not only explain how to meet the quality criteria of validity, reliability, transparency and independence but also include the conditions that increase the likelihood of meeting these criteria, such as the importance of examiners' assessment expertise and how the institution should facilitate their development in this area. The case study demonstrates how these guidelines are applied to examine thesis assessment practices at a bachelor's psychology-related program at a Dutch academic university.

Our experience highlights the importance of applying the didactic principle of constructive alignment at the exit level, as it is not always clear to teaching staff what this means in the context of thesis assessment (despite its widespread use at the course level for instructional design) and how it can be used to ensure the four standards of the Framework. This has led to a focus on reliability, as noted by Webster *et al.* (2000), such as revising thesis assessment forms and ensuring consistency among examiners. Our study aims to draw the attention of program teams to validity by considering the program's curriculum and assessment design and the didactic purpose of using a thesis as a graduation project.

While other studies have focused on specific thesis assessment quality criteria such as reliability (e.g. Pathirage *et al.*, 2007), transparency (e.g. Malcolm, 2020) and independence (Todd *et al.*, 2004; e.g. Nyamapfene, 2012), our case study shows how to ensure all of these criteria and carry out a complete quality assurance process. This does not mean that a program needs to address all of them at the same time. Instead, we want to emphasize the importance of research education in a bachelor's program and recommend that the program align its thesis assessment design with its curriculum design for research education (i.e. as a learning trajectory) and its overall assessment design. Improving thesis assessment alone is not sufficient for students to achieve the intended learning outcomes of the program.

A final, and perhaps the most important, aspect to consider is how to effectively use limited resources to improve teaching staff's assessment expertise so that they can continuously contribute to the improvement of thesis assessment practices. The guidelines presented in this study can be further developed or adapted as training materials for teaching staff.

Limitations

We would like to acknowledge two limitations of this study. First, unlike more traditional research methods such as surveys and interviews, the problems we reported here were compiled from various sources at four Dutch research universities. Without a more rigorous synthesis of these sources, it is possible that there may be some subjectivity and selection bias present. Second, the guidelines we derived from a narrative review of these problem topics may not include all relevant references.

It is important to note that our use of only one psychology-related bachelor's program for the case study does not allow us to generalize our findings to all bachelor's psychology programs at other Dutch academic universities. Rather, our aim is to share our experience and research-informed guidelines, and to examine thesis assessment quality from a practitioner perspective. In line with the goals of Koris and Pello's (2022) article, our aim is to gradually find solutions that are appropriate for our context through several subsequent iterations in the future.

Notes

1. https://www.universiteitenvannederland.nl/en_GB/f_c_ingeschreven_studenten.html
2. https://www.universiteitenvannederland.nl/en_GB/reduce-work-pressure#eerste

References

Andriessen, D. and Manders, P. (2013), *Beoordelen Is Mensenwerk [Evaluation Is Human Work]*, Vereniging Hogescholen, Den Haag.

Bamber, M. (2015), "The impact on stakeholder confidence of increased transparency in the examination assessment process", *Assessment and Evaluation in Higher Education*, Vol. 40, pp. 471-487.

Bell, A., Mladenovic, R. and Price, M. (2013), "Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars", *Assessment and Evaluation in Higher Education*, Vol. 38, pp. 769-788.

Bergwerff, M. and Klaren, M. (2016), *Kwaliteitsborging Toetsing: En Handreiking Voor Examencommissies [Quality Assurance of Assessment: A Guide for Examination Boards]*, Leiden University, Leiden.

Bettany-Saltikov, J., Kilinc, S. and Stow, K. (2009), "Bones, boys, bombs and booze: an exploratory study of the reliability of marking dissertations across disciplines", *Assessment and Evaluation in Higher Education*, Vol. 34, pp. 621-639.

Biggs, J. and Tang, C. (2007), *Teaching for Quality Learning at University*, Open University Press, New York, NY.

Bloxham, S. and Boyd, P. (2007), *Developing Effective Assessment in Higher Education: A Practical Guide*, McGraw-Hill Education, New York, NY.

Bloxham, S., Boyd, P. and Orr, S. (2011), "Mark my words: the role of assessment criteria in UK higher education grading practices", *Studies in Higher Education*, Vol. 36, pp. 655-670.

Bloxham, S., den-Outer, B., Hudson, J. and Price, M. (2016a), "Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria", *Assessment and Evaluation in Higher Education*, Vol. 41, pp. 466-481.

Bloxham, S., Hughes, C. and Adie, L. (2016b), "What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices", *Assessment and Evaluation in Higher Education*, Vol. 41, pp. 638-653.

Bologna Working Group (2005), *A Framework for Qualifications of the European Higher Education Area. Bologna Working Group Report on Qualifications Frameworks*, Danish Ministry of Science, Technology and Innovation, Copenhagen.

Brookhart, S.M. (2013), *How to Create and Use Rubrics for Formative Assessment and Grading*, Association for Supervision and Curriculum Development (ASCD), Alexandria, VA.

Brookhart, S.M. (2018), "Appropriate criteria: key to effective rubrics", *Frontiers in Education*, [Online], Vol. 3, available at: https://www.frontiersin.org/article/10.3389/feduc.2018.00022 (accessed 10 April 2018).

Dierick, S., van de Watering, G.A. and Muijtjens, A. (2002), "De actuele kwaliteit van assessment: ontwikkelingen in de edumetrie", in Dochy, F., Heylen, L. and van de Mosselaer, H. (Eds), *Assessment in onderwijs: Nieuwe toetsvormen en examinering in studentgericht onderwijs en competentiegericht onderwijs*, Boom Lemma Uitgevers, Amsterdam.

Golding, C., Sharmini, S. and Lazarovitch, A. (2014), "What examiners do: what thesis students should know", *Assessment and Evaluation in Higher Education*, Vol. 39, pp. 563-576.

Hand, L. and Clewes, D. (2000), "Marking the difference: an investigation of the criteria used for assessing undergraduate dissertations in a business school", *Assessment and Evaluation in Higher Education*, Vol. 25, pp. 5-21.

Hsiao, Y.P. and Verhagen, M. (2018), "Examining the quality and use of the grading form to assess undergraduate theses", *9th Biennial Conference of EARLI SIG 1: Assessment and Evaluation*, Helsinki, Finland.

Inspectorate of Education [Inspectie van het Onderwijs] (2016), *De kwaliteit van de toetsing in het hoger onderwijs [The assessment quality in higher education]*, Ministry of Education, Culture and Science [Ministerie van Onderwijs, Cultuur en Wetenschap], Utrecht.

Johnson, R.L., Penny, J., Gordon, B., Shumate, S.R. and Fisher, S.P. (2005), "Resolving score differences in the rating of writing samples: does discussion improve the accuracy of scores?", *Language Assessment Quarterly*, Vol. 2, pp. 117-146.

Kiley, M. and Mullins, G. (2004), "Examining the examiners: how inexperienced examiners approach the assessment of research theses", *International Journal of Educational Research*, Vol. 41, pp. 121-135.

Koris, R. and Pello, R. (2022), "We cannot agree to disagree: ensuring consistency, transparency and fairness across bachelor thesis writing, supervision and evaluation", *Assessment & Evaluation in Higher Education*, pp. 1-12.

Malcolm, M. (2020), "The challenge of achieving transparency in undergraduate honours-level dissertation supervision", *Teaching in Higher Education*, Vol. 28 No. 1, pp. 1-17.

Maxwell, T.W. and Smyth, R. (2011), "Higher degree research supervision: from practice toward theory", *Higher Education Research and Development*, Vol. 30, pp. 219-231.

McQuade, R., Kometa, S., Brown, J., Bevitt, D. and Hall, J. (2020), "Research project assessments and supervisor marking: maintaining academic rigour through robust reconciliation processes", *Assessment and Evaluation in Higher Education*, Vol. 45, pp. 1181-1191.

Mullins, G. and Kiley, M. (2002), "'It's a PhD, not a Nobel Prize': how experienced examiners assess research theses", *Studies in Higher Education*, Vol. 27, pp. 369-386.

NLQF (2008), *The Higher Education Qualifications Framework in the Netherlands, a Presentation for Compatibility with the Framework for Qualifications of the European Higher Education Area*, Dutch Qualifications Framework (NLQF), 's-Hertogenbosch, The Netherlands.

NVAO (2008), *The Higher Education Qualifications Framework in the Netherlands: A Presentation for Compatibility with the Framework for Qualifications of the European Higher Education Area*, Accreditation Organisation of the Netherlands and Flanders (NVAO), Den Haag.

NVAO (2018), *Assessment Framework for the Higher Education Accreditation System of the Netherlands*, Accreditation Organisation of the Netherlands and Flanders (NVAO), Den Haag.

Nyamapfene, A. (2012), "Involving supervisors in assessing undergraduate student projects: is double marking robust?", *Engineering Education*, Vol. 7, pp. 40-47.

O'Donovan, B., Price, M. and Rust, C. (2004), "Know what I mean? Enhancing student understanding of assessment standards and criteria", *Teaching in Higher Education*, Vol. 9, pp. 325-335.

Orsmond, P., Merry, S. and Reiling, K. (2002), "The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment", *Assessment and Evaluation in Higher Education*, Vol. 27, pp. 309-323.

Osborn Popp, S.E., Ryan, J.M. and Thompson, M.S. (2009), "The critical role of anchor paper selection in writing assessment", *Applied Measurement in Education*, Vol. 22, pp. 255-271.

Pathirage, C., Haigh, R., Amaratunga, D. and Baldry, D. (2007), "Enhancing the quality and consistency of undergraduate dissertation assessment: a case study", *Quality Assurance in Education*, Vol. 15, pp. 271-286.

Price, M. (2005), "Assessment standards: the role of communities of practice and the scholarship of assessment", *Assessment and Evaluation in Higher Education*, Vol. 30, pp. 215-230.

Reguant, M., Martínez-Olmo, F. and Contreras-Higuera, W. (2018), "Supervisors' perceptions of research competencies in the final-year project", *Educational Research*, Vol. 60, pp. 113-129.

Rust, C., Price, M. and O'Donovan, B. (2003), "Improving students' learning by developing their understanding of assessment criteria and processes", *Assessment and Evaluation in Higher Education*, Vol. 28, pp. 147-164.

Sadler, D.R. (1987), "Specifying and promulgating achievement standards", *Oxford Review of Education*, Vol. 13, pp. 191-209.

Sadler, D.R. (2013), "Assuring academic achievement standards: from moderation to calibration", *Assessment in Education: Principles, Policy and Practice*, Vol. 20, pp. 5-19.

Shay, S. (2005), "The assessment of complex tasks: a double reading", *Studies in Higher Education*, Vol. 30, pp. 663-679.

Todd, M., Bannister, P. and Clegg, S. (2004), "Independent inquiry and the undergraduate dissertation: perceptions and experiences of final-year social science students", *Assessment and Evaluation in Higher Education*, Vol. 29, pp. 335-355.

University of Twente (2019), *The Bachelor's Thesis [Online]*. University of Twente, Enschede, The Netherlands, available at: https://www.utwente.nl/en/bee/programme/thesis/ (accessed 8 September 2019).

Walvoord, B.E. and Anderson, V.J. (2011), *Effective Grading: A Tool for Learning and Assessment in College*, Jossey-Bass, San Francisco, CA.

Webster, F., Pepper, D. and Jenkins, A. (2000), "Assessing the undergraduate dissertation", *Assessment and Evaluation in Higher Education*, Vol. 25, pp. 71-80.

Wijngaards-de Meij, L. and Merx, S. (2018), "Improving curriculum alignment and achieving learning goals by making the curriculum visible", *International Journal for Academic Development*, Vol. 23, pp. 219-231.

Willison, J.W. (2012), "When academics integrate research skill development in the curriculum", *Higher Education Research and Development*, Vol. 31, pp. 905-919.

Willison, J.W. and O'Regan, K. (2006), *Research Skill Development Framework [Online]*, The University of Adelaide, Adelaide, available at: https://www.adelaide.edu.au/rsd/ (accessed 19th November 2019).

Wylie, E.C. and Szpara, M.Y. (2004), *National Board for Professional Teaching Standards Bias-Reduction Training: Impact on Assessors' Awareness*, ETS Research Report Series, Princeton, NJ, Vol. 2004, p. i-55.

About the authors

Ya-Ping (Amy) Hsiao is an assessment specialist and teacher trainer at Tilburg University. Her current research focuses on the reflection, portfolio and performance assessment of the graduation projects. Ya-Ping (Amy) Hsiao is the corresponding author and can be contacted at: y.p.hsiao@tilburguniversity.edu

Gerard van de Watering is a policy advisor at Eindhoven University of Technology. His research and development interest focus on assessment and evaluation, student-centered learning environments,

independent learning and study skills. He is also the founder of a network of assessment specialists in academic higher education in the Netherlands.

Marthe Heitbrink is a testing and assessment coordinator at the Psychology department of the University of Amsterdam.

Helma Vlas is an educational consultant, teacher trainer/assessor and assessment specialist at the University of Twente. She is stationed at the Centre of Expertise in Learning and Teaching. She is coordinator of the Senior Examination Qualification trajectory at the University of Twente.

Mei-Shiu Chiu is a full professor of Education at National Chengchi University in Taiwan. Her research interests focus on interactions between emotion/affect, cognition and culture for diverse knowledge domains (e.g. mathematics, science and energy) in relation to teaching, assessment and large-scale databases.