Assessing student theses: Differences and similarities between examiners from different academic disciplines

Practitioner Research in Higher Education
Special Assessment Issue

University of Cumbria
Vol 10(1) pages 217-226

Mats Lundström[1], Maria Åström[2], Karin Stolpe[3], Lasse Björklund[3]
1 Malmö University, 2 Umeå University, 3 Linköping University
mats.lundstrom@mah.se

Abstract
The writing of student theses is an important activity at universities and is expected to demonstrate the students' academic skills. In the teacher-education programme, examiners from different academic disciplines are involved in supervising and examining student theses. Moreover, different subject disciplines have different traditions concerning what is seen as knowledge and the way research is performed, which could result in different assessment practices and judgements. Earlier studies demonstrate a fragmented picture concerning the importance of the examiners' academic discipline in judging theses. The purpose of this article is to investigate whether examiners from different academic subject disciplines emphasise similar or different criteria when assessing student theses. A total of 66 examiners from six universities with teacher education programmes in Sweden have answered an online Q-survey where they compared different criteria and rank-ordered them. The results demonstrate minor differences between individuals from different academic disciplines: Only two out of the 45 criteria had significant differences between academic discipline groups. Thus, the results indicate that teacher education is a boundary-crossing, multi-disciplinary field which primarily uses generic criteria.

Keywords
Assessment criteria; assessment culture; judgement, student theses; repertory grid technique.

Introduction
Academic writing is a significant activity at universities, and writing a thesis is an established activity which is expected to demonstrate a student's academic skills. During the past year, Swedish teacher education has focused on student theses to an even higher degree than before; this is justified in terms of the desire for a more scientific teacher education (Högskoleverket, 2010; SOU, 2008). Teacher education in Sweden is 3.5-5 years, and it is offered at approximately 25 universities or college universities. Student theses are important both in the student-teachers' examination and in the evaluation of different teacher programmes. Further, the importance of writing a thesis as regards both a future career as a teacher and as a researcher have been considerably discussed (Råde, 2014). In earlier teacher-education programmes, student teachers wrote one thesis, worth 15 credits (10 weeks of studies), at the end of their education. Since the last teacher-education reform (2011), a majority of teaching students have been writing theses equivalent to 30 credits at the end of their education, which can be done either as one large thesis or, more commonly, as two different theses. It is up to the university to decide which of these two approaches is applied. One common model in Sweden is that the student first writes a thesis which is a systematic review, whereby earlier research in a field is summarised and analysed. At a later stage, the student writes a second thesis on a more advanced level. This second thesis is more empirical in that the student is required to collect data, for example, through observations, surveys or interviews, moreover, the second thesis qualifies the student to apply to a PhD programme.

Citation
Lundström, M., Åström, M., Stolpe K., Björklund L. (2016) 'Assessing student theses: Differences and similarities between examiners from different academic disciplines', *Practitioner Research in Higher Education Journal*, 10(1), 217-226.

This increasing number of student theses written in teacher-education programmes necessitates more supervisors and examiners. Traditionally, teacher education in Sweden consists of supervisors and examiners from different academic disciplines: for example, pedagogy, science, social science, psychology, sociology, mathematics and modern languages. The mix of academics from different disciplines has risen even more since the last teacher-education reform. These different subject disciplines have different traditions concerning what is seen as knowledge and the way research is performed (Schwab, 1964; Scriven, 1964). Moreover, it is worth noting that while some of supervisors and examiners have, themselves, graduated from a teacher-education programme, others have not. The differences between disciplines might result in diverse views concerning what constitutes a good student thesis, something which could undermine their trustworthiness among students. Consequently, since student theses are seen as an important factor in measuring the quality of an education, it is fundamental that all assessors and examiners of student theses have a similar view of the assessment criteria. Therefore, the purpose of this article is to investigate whether examiners from different academic disciplines emphasise similar or different criteria when assessing student theses.

Earlier research concerning assessment of student theses

An examiner's disciplinary background is often regarded as an important factor in the examination of theses. Becher (1994) describes disciplinary differences with different cultural characteristics, building his argumentation on 350 interviews with respondents from 12 different disciplines (biology, chemistry, economy, engineering, geography, history, law, mathematics, modern languages, pharmacology, physics and sociology). These fields have common points of contact academically; however, they have different traditions regarding knowledge and, by extension, how examination is performed. Moreover, Woolf (2004) states that different academic departments have different cultures that influence the marking process. He conducted a small-scale investigation into the criteria used by a number of different departments for assessing final-year project modules in business and history as well as other written history assignments. However, when investigating the marking criteria, Woolf (2004) found that they might not be distinctly subject-specific in some departments, and he suggests that this could be explained by the fact that, in these departments, generic criteria are viewed as the best way to judge a thesis. Similarly, other studies have revealed small differences in how examiners judge theses; for example, Bettany-Saltikov, Kilinc and Stow (2009) found high inter-reliability between examiners from different disciplines. The examiners in their study assessed master's-level essays and used general, overarching criteria based on academic practices that all in the profession might recognize: use of evidence; technical referencing; clear, logical communication and relevance. However, with only four respondents, the study was limited. Similarly, Kiley and Mullins (2004) did not find any differences in examination practice that could be explained by disciplinary background. Even when the examiners' marking process reflected their disciplinary background, it did not result in any differences in judgement.

Some studies give a mixed picture of the importance of disciplinary background. Ekbrand et al. (2014) conducted a study based on written evaluations which were provided by graders as part of the examination process of theses from a teacher-education programme in Sweden. The results show that inter-examiner differences are substantial and much greater than both inter-departmental differences and inter-faculty differences. However, there are substantial inter-faculty differences when it comes to what the examiners gave the students in terms of imperatives, questions and explanations. As regards criteria-codes, examiners from the sciences stood out as more interested in questions about analysis compared to examiners from the social sciences or the humanities; on the other hand, the preoccupation with methodology and aim was characteristic of the social sciences. The authors draw the conclusion that the subcultures of departments or faculties do not significantly impact the assessment practice. Although, the differences between examiners in the same department overshadow inter-departmental differences.

Some studies are more in line with Becher's (1994) view of cultural characteristics and have reported differences between examiners from different traditions. A study by Erixon (2011) explores the perceptions of researchers from different scientific and scholarly areas about scientific and scholarly writing. The study included interviews with 12 researchers in four different faculties – Arts, Social Sciences, Science and Technology, and Medicine – at a Swedish university. Erixon (2011) uses an analytical tool based on Biglan's (1973) and Becher's (1994) four intellectual clusters: hard pure (natural) science, soft pure (arts and social) sciences, hard applied (engineering) sciences and soft applied (education) sciences. The findings suggest that researchers in the applied sciences regard writing as having a mediating and creative function for research, while pure scientists view writing as based on epistemology that does not attribute a mediating function to language (Wertsch, 1998). The study also indicates that applied-science researchers (i.e., those with professional education of various kinds) are positioned at the interface between disciplines and individuals as social beings and that they operate as epistemological boundary-crossers for the faculties (Erixon, 2011).

Suto and Greatorex (2008) have demonstrated how examiners from different disciplines (business and mathematics) employ different strategies when they assess written examinations. Examiners with mathematics backgrounds used diverse strategies to a greater extent, mixing intuitive methods with more evaluative ones. This might be explained by the differences between business and mathematics as fields, but it also raises the question of whether these strategies are used also when examining a student thesis since a student thesis in a teacher programme can represent a combination of different subjects, for instance, pedagogy and science. Similarly, Erixon Arreman and Erixon (2015) have found differences between researchers from different disciplines regarding what they see as a good student thesis. The respondents in this interview study often considered the writing traditions from their own academic disciplines when evaluating the structure of a student thesis in the teacher-education programme.

Research about assessment is not only a matter of what examiners express as important but also a matter of how these expressions get into practice. Bloxham et al. (2015) have demonstrated how experienced examiners from different disciplines judge student work. The 24 examiners from four different disciplines were provided with relevant assessment criteria and five examples of student work from each discipline. Through the use of the repertory grid method, Bloxham et al. found that only one of the 20 assignments was assigned the same rank by all six assessors from the same discipline. All the other assignments were given grades that 'ranked' them against the other assignments in at least three different positions (i.e., best, second best and so on). Nine of the 20 assignments were ranked both best and worst by different assessors. Assuming that the constructs represent examiners' implicit criteria, Bloxham et al. draw the conclusion that although they appear to use similar criteria, in practice, the examiners interpret such criteria differently, and this has the potential to contribute to differences in standards. Moreover, Bloxham et al. conclude that assessors have different expectations as regards the standards required at various levels.

To summarise, earlier studies present a divided view on the importance of examiners' disciplines when assessing theses. On the one hand, a different view of knowledge and culture (Becher, 1994; Woolf, 2004) might result in a different view and practice (Erixon, 2011; Erixon Arreman and Erixon, 2015) or assessment strategy (Suto and Greatorex, 2008). On the other hand, some studies have indicated that differences in how examiners perform assessments are insignificant or non-existent (Bettany-Saltikov et al., 2009; Kiley and Mullins, 2004). Furthermore, some studies have indicated that even when there are some differences, they are not always the result of academic background or faculty affiliation (e.g., Ekbrand, et al., 2014). Bloxham et al. (2015) explain the differences between examiners in terms of different interpretations of criteria and different expectations of required standards. These contradictory results highlight the need for more studies – studies that investigate disciplinary background and view of criteria. The use of criteria is common at many

universities and is part of the practice among examiners at these universities, which makes the focus on criteria even more relevant.

Research questions

This study poses the following questions: Which criteria are emphasised among examiners from different academic disciplines when examining student theses?
How can the similarities and differences between the examiners be explained?

Method

Data was collected through an online Q-survey used in Q-methodology (Brown, 1997; Stephenson, 1953). In a Q-survey, the informant has to take a stance on a number of different criteria and rank-order them. The criteria in this Q-survey were generated through interviews with 19 respondents who have experience as examiners of student theses. The respondents were from three different universities in Sweden and chosen to represent different experiences as examiners and different PhD subjects. The interviews were conducted a couple of months before the Q-survey, and the purpose of these interviews was to identify the criteria used by examiners in teacher-education programmes. The interviews were individual and carried out using a combination of Comparative Judgement (Pollit, 2012; Thurstone, 1927) and Repertory Grid Technique (RGT) (Björklund, 2008; Kelly, 1955). In the Comparative Judgement, five to seven student theses, written by undergraduate students, were compared pair-wise. Comparative judgment was proposed by Louis Thurstone as a method for constructing a ranking-scale based on direct comparisons of pairs of objects. Pollit et al. (2012) introduced this method to educational assessment in England in 1993, and it has since become the regular experimental method for comparing standards.

In the RGT interviews, the respondents received the task of choosing five to seven student theses (same as above) that they had read before the interview (they could choose theses which they had supervised or examined). During the interview, the interviewer randomly picked three of the student theses and asked the informant to pick one that differed from the other two. Thereafter, the interviewee was questioned about the way in which this particular thesis differed from the other two. This difference was then established as a construct/criterion. The methodology of the project is more accurately described in Björklund, Stolpe and Lundström (2016). During the interviews, 92 different criteria were mentioned. Several of them were similar, although not exactly the same. Two of the article's authors then categorised and reduced the number of different criteria to 52. The criteria were mainly traditional ones which can be found in many universities' assessment guidelines or matrices. The criteria that were similar, but not exactly the same, were reduced or merged. For example, *scientific foundation* and *good research anchoring* were merged into *research anchoring*. The criteria *good literature* was reduced since *relevant literature* already was a criterion. In the final construction of the Q-survey, seven more criteria were reduced or merged in a similar way to make the survey more manageable for the respondents. This done in the same way as the earlier criteria reduction. For instance, *discussion about research ethics* was reduced since *research ethics* was already one of the criteria. The authors then categorised the remaining 45 criteria into main categories: *relation to research, definition of research scope, theory, method, language and formalities, performance and conclusions, general totality* and *miscellany*. These main categories have been used in earlier studies concerning student theses (Högskoleverket, 2006; Råde, 2014). The 45 criteria were analysed in another part of the project and compared with the official criteria of the universities involved in the study; this official criteria can be described as generic criteria (Lundström et al., submitted).

The Q-survey was divided into two steps. In the first step, the respondents had to sort the 45 criteria into three piles: very important, important and less important. In the next step, they had to sort the same criteria further: from -5 to +5 on an 11-grade scale. It was not possible for them to rank all criteria highly; they were forced to spread the criteria evenly below a Gaussian border. This forced

the respondents to really consider which criteria they found most important and to always compare with the neighbouring criteria. A pilot survey was constructed and tested on nine respondents. Since this pilot study did not lead to any questions from the respondents, no changes were made to the survey.

The survey was conducted online and sent out to 179 respondents, including those from the RGT interviews, at six different universities in Sweden. The majority of the respondents came from three different universities that have a large number of teaching students, and all the respondents were recognised as examiners of student theses. After three weeks, a reminder was sent out to those who had not answered the survey. In the end, 66 respondents had answered the Q-survey, which gives a respondent rate of 36.9%. A loss analysis of the individuals who did not answer the survey was made. Background variables such as sex, university, experience as tutor and examiner, teacher education and academic discipline (PhD subject) were also collected through the survey. None of these variables were over-represented in the analysis of loss. The examiners were divided into three groups based on the subject of their PhD: group 1, science, mathematics or psychology; group 2, pedagogy or other educational science; and group 3, any other subject (e.g., history or ethnology). Four of the respondents did not have a PhD degree. The respondents had different experience as examiners, from one year of experience supervising or examining less than 10 student theses to over 20 years of experience supervising or examining over 100 student theses.

Ethical approval

The interviews and the online survey were voluntary. The identities of all the respondents, both from interviews and the survey, are confidential; only the four researchers conducting the project know the respondents' identities. All the examiners interviewed with RGT signed a consent document.

Analysis

Data from the Q-survey was analysed using traditional measure theory using SPSS (Statistical Package for the Social Sciences). The results from the Q-analysis are also described and reported in Stolpe et al. (in preparation). The scale in the initial survey (-5 to + 5) was reconstructed and translated to a scale of 0 to 10 to make calculations easier. Mean values, standard deviations and significances were calculated both for each of the 45 criteria and for the eight main categories involving more than one criterion (Edling and Hedström, 2003; Robson, 2002). The main categories include various numbers of criteria (3–8). In order to investigate whether some particular criterion or main category was more important for examiners from a particular academic discipline, the mean value, standard deviance and significance of all 45 criteria were analysed for each of the groups described above through analysis of variance (ANOVA).

The majority of the criteria can be described as generic criteria (Woolf, 2004). This is in line with Erixon's (2011) description of professional education as an epistemological boundary crosser for the faculties.

Results

In Table 1, the means values and standard deviations of the eight main categories are presented.

Table 1. Main categories' mean values and standard deviations.

Main category	Mean	Std dev
Relation to research	3.68	0.68
Definition of research scope	6.29	1.26
Theory	7.08	1.39
Method	5.55	1.03
Language and formalities	4.82	1.16
Performance and conclusions	7.37	0.82
General totality	3.88	0.61
Miscellany	3.60	0.76

As we can see in Table 1, the traditional main categories – *performance and conclusions*, *theory* and *definition of research scope* – are ranked as most important for examiners according to the mean values. However, it is less important to *relate to research* or to have correct *language and formalities*, which are also traditional categories in assessing theses. *General totality* and *miscellany* are ranked low according to the mean value.

Table 2. Mean values and standard deviation for three groups of examiners from different academic disciplines.

Group		Group 1 N=10	Group 2 N=23	Group 3 N=29	Total N=62
Relation to research	Mean	3.86	3.85	3.49	3.68
	Std dev	0.53	0.83	0.57	0.69
Definition of research scope	Mean	6.07	5.88	6.59	6.25
	Std dev	1.18	1.43	1.06	1.26
Theory	Mean	7.35	6.74	7.17	7.04
	Std dev	1.11	1.36	1.51	1.40
Method	Mean	5.24	5.87	5.54	5.62
	Std dev	1.03	1.27	0.78	1.03
Language and formalities	Mean	4.28	4.85	5.02	4.82
	Std dev	0.81	1.38	1.12	1.20
Performance and conclusions	Mean	8.00	7.02	7.38	7.35
	Std dev	0.68	0.90	0.70	0.83
General totality	Mean	3.80	4.03	3.79	3.88
	Std dev	0.79	0.59	0.58	0.62
Miscellany	Mean	3.55	3.67	3.57	3.60
	Std dev	0.59	0.81	0.80	0.77

Means and standard deviations for both criteria and categories were calculated with the help of SPSS. An analysis of variance was made to investigate whether any differences between the mean values between the groups were significant. If we compare how individuals from different academic disciplines answered, there is only one significant result (p= 0.006) where academic discipline matters (i.e., if the examiner was in group 1, 2 or 3 described above). The group comprised of individuals who have a PhD degree in mathematics, science or psychology (group 1) emphasises *performance and conclusions* compared to individuals with degrees in other academic disciplines (group 2 and group 3) as shown in Table 2. The mean value for the math/sci/psy-group (group 1) was 8.00 compared to a mean of 7.02 in group 2 and a mean value of 7.38 in group 3 for this category. No significant differences were noticed in the analysis of the rest of the main categories. This result

indicates that examiners from different academic disciplines emphasise similar parts of a student thesis. The rather low standard deviations indicate small differences within and between the groups. The main categories include various numbers of criteria (3-8). To investigate whether some criteria were more important for examiners from particular academic disciplines, the means of all 45 criteria were analysed for all three groups. The analysis indicates insignificant differences between examiners from different disciplines. In only two of the 45 criteria could any significant differences ($p \leq 0.05$) be found. The group of individuals that have a PhD degree in pedagogy and subject didactics (group 2) emphasise *transparent method*. *Substantiated conclusions* are important for the math/sci/psy group (group 1) and for those who do not have a PhD. These results indicate that academic discipline does not play a significant role in the selection of criteria examiners think are important in a student thesis.

The majority of the eight main categories in Table 1 consist of traditional criteria such as *connections to research*, *researchable purpose*, *use of theory* and *appropriate choice of method*. There are two main categories which are different and consist of more non-traditional criteria: *general totality* and *miscellany*. As examples of criteria from the category *general totality*, we can take *low degree of normativity* and *gender perspective*, and from the category *miscellany*, we can take *exciting and strong narrative voice*, *low degree of normativity* and *gender perspective*. If we look at Table 1, we can see that both *general totality* and *miscellany* are ranked low (3.88 and 3.60). This result indicates traditional criteria as more important than non-traditional ones and applies to all three groups in the study.

One criterion is strongly related to student teachers' future careers. The criterion vocational relevance investigates whether the examiners see the involvement of upcoming teacher practice as important in the student thesis. The mean value for the criterion was 4.63, and the standard deviation was 2.24, which demonstrates that, according to the respondents, vocational relevance is not so important in a student thesis.

The data from the Q-survey has also been analysed to investigate differences which can be explained by the examiner's experience or university. No significant differences could be detected. These results will be presented in Stolpe et al. (in preparation).

Discussion

Our results demonstrate small differences between examiners from different disciplines. Significant differences are detected only in one of eight main categories and two of 45 criteria, which can be regarded as minor differences between examiners. On the one hand, academic discipline has often been viewed as important in its own assessment practice (e.g., Becher, 1994; Erixon, 2011; Erixon and Arreman, 2015; Erixon, 2015; Suto and Greatorex, 2008; Woolf, 2004), which should be related to and explained by culture and traditions within different academic fields. On the other hand, Kiley and Mullins (2004) found no examination-practice differences that could be explained with disciplinary background, which is more similar to the results in our study. Our results are between these two extremes, with only minor differences when we compare examiners from different academic disciplines. The main difference we found (in emphasising *performance and conclusions* among examiners from the sciences) is in line with the results of Ekbrand et al. (2014), who found that examiners from the sciences stand out as more interested in questions about analysis. *Analysis* was, in our study, one of the criteria in the main category *performance and conclusions*. All three groups in our study emphasised traditional criteria more than non-traditional ones, demonstrating that none of the three groups representing different academic disciplines depart from what is traditionally regarded as important in a student thesis.

The examiners' experience – how long they have worked at universities with teacher-education programmes and how many student theses they have examined – differs markedly. Although they

have different experience levels, one explanation of our results could be that a common understanding and assessment practice will quickly evolve in a university with a teacher-education programme. However, the respondents in the study came from different universities, and even within a university, several departments are involved in the student theses. Ekbrand et al. (2014) have demonstrated how differences between examiners in the same department overshadow the differences between departments. Similarly, it was not possible to detect significant differences between departments or universities in our study (Stolpe et al., in preparation).

The small differences between examiners from different academic disciplines might be explained by the fact that the main categories and criteria can be regarded as general and not subject-specific. Woolf (2004) and Bettany-Saltikov, Kilinc and Stow (2009) have earlier demonstrated small differences in assessment between examiners from different academic fields when the criteria are general. Erixon's (2011) description of educational sciences as cross-boundary and positioned at the interface between other disciplines may lead to a view whereby general or more overarching criteria are used. A preliminary analysis of the official criteria at the different universities represented in this study show that general criteria appear to be common at Swedish universities with teacher-education programmes. This might lead to small differences in emphasised criteria between examiners at a teacher-education programme. Despite the fact that the students are becoming teachers, the criterion vocational experience is ranked low, although there is some disagreement. The majority of the examiners regard the student thesis as a general academic performance. According to them, the connection to future work is not so important.

One weakness of this study is that answering a query is not the same as assessing theses in real life. However, similar to Bloxham et al. (2015), we contend that ranking data can be considered to provide a broad picture of examiners' views. Nevertheless, this study indicates that academic discipline is not the most important factor in explaining differences in examiners' assessment practices; rather, the differences which have been reported (Erixon and Arreman, 2015; Erixon, 2015; Suto and Greatorex, 2008) might be explained in other ways. Bloxham et al. (2015) suggest that although assessors appear to use similar criteria, in practice, they interpret such criteria differently; this has the potential to contribute to differences in standards. They mean that a personalized *standard framework* (Bloxham, Boyd and Orr, 2011) is developed in the individuals who engage in reading student work. This standard framework is dynamic and influenced, but it is not determined by subject discipline norms (Shay, 2005). We think this suggestion about an individual standard framework is interesting, and it erases some of the differences between different academic disciplines concerning what is regarded as important. If we combine the individual standard framework (Bloxham, Boyd and Orr., 2011) with Erixon's (2011) suggestion about professional education as an interface between the disciplines and as epistemological boundary-crossers for the faculties, an explanation to the small differences between different groups in our study could be provided. We suggest further studies that investigate examiners' explicit criteria during assessment. One such study has been reported in this project (Björklund, Stolpe and Lundström, 2016, in which examiners ranked student theses using a combination of Comparative Judgement (Pollit, 2012; Thurstone, 1927) and Repertory Grid Technique (RGT) (Björklund, 2008; Kelly, 1955). Another possible study would be to examine similarities and differences in views between supervisors and examiners of the same student thesis and in this way investigate how the supervising process affects judgement of a student thesis.

Acknowledgements
This research project has been funded by the Swedish Research Council.

References

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher Education. , 19* (2), 151-162.

Bettany-Saltikova, J., Kilinc, S., & Stowc, K. (2009). Bones, boys, bombs and booze: an exploratory study of the reliability of marking dissertations across disciplines. *Assessment & Evaluation in Higher Education, 34* (6), 621–639.

Biglan, A. (1973). The characteristics of subject matter in different scientific areas. *Journal of Applied Psychology 57*, 195-203.

Björklund, L. (2008b). The Repertory Grid Technique: Making Tacit Knowledge Explicit: Assessing Creative work and Problem solving skills. In H. Middleton (Ed.), *Researching Technology Education: Methods and techniques.* Netherlands,: Sense Publishers.

Björklund, L., Lundström, M. & Stolpe, K. (2016). Making tacit knowledge explicit. Three methods to assess attitudes and believes. Published in J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto & K. Hahl. Science Education Research: Engaging learners for a sustainable future. *E-proceeding from ESERA*, 2015.

Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education, 36*(6), 655-670.

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2015). Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria,. *Assessment & Evaluation in Higher Education*. doi: 10.1080/02602938.2015.1024607

Brown, S. R. (1997). The History and Principles of Q Methodology in Psychology and the Social Sciences. Retrieved 090115 from http://facstaff.uww.edu/cottlec/QArchive/Bps.htm

Edling, C., & Hedström, P. (2003). *Kvantitativa metoder. Grundläggande analysmetoder för samhälls- och beteendevetare*. Lund: Studentlitteratur.

Ekbrand, H., Gunnarsson, A., Jansson, M., & Carle, J. (2014). *Assessment in higher education: exploring the gap between standards and idiosyncrasies.* Paper presented at the Sociologidagarna, Göteborgs universitet.

Erixon Arreman, I., & Erixon, P.-O. (2015). The degree project in Swedish Early Childhood Education and Care - what is at stake? *Education Inquiry, 6*(3), 309-332.

Erixon, P.-O. (2011). Academic Literacies; Discourse and Epistemology in a Swedish University. *Education Inquiry, 2*(2), 221-238.

Högskoleverket. (2006). Examensarbetet inom den nya lärarutbildningen. Stockholm.

Högskoleverket. (2010). Högskoleverkets system för kvalitetsutvärdering 2011–2014. In Högskoleverket (Ed.). Stockholm.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Routledge.

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

Lundström, M., Åström,M., Stolpe, K., & Björklund, L. (submitted). Konsumtionsuppsatser som ny praktik för lärarutbildare. [Student theses as new practice for teacher educators)

Pollit, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*, 157-170.

Robson, C. (2002). *Real World Research. A Resourse for Social Scientists and Practitioner-Researchers* (2nd ed.). Cornwall, UK: Blackwell Publishing.

Råde, A. (2014). Ett examensarbete för både yrke och akademi – En utmaning för lärarutbildningen. *Högre utbildning, 4*(1), 19-34.

Schwab, J. J. (1964). The structure of the Natural Sciences. In J. C. Parker (Ed.), *The structure of Knowledge and the Curriculum* (Second ed., pp. 31-49). U.S.A: Rand McNally & Company.

Scriven, M. (1964). The structure of the Social Studies. In J. C. Parker (Ed.), *The structure of Knowledge and the Curriculum* (Second ed., pp. 87-105). U.S.A: Rand McNally & Company.

Shay, S. B. (2005). The assessment of complex tasks: a double reading. *Studies in Higher Education, 30*(6), 663-679. doi: 10.1080/03075070500339988

SOU. (2008). *En hållbar lärarutbildning. Betänkande av Utredningen om en ny lärarutbildning (HUT 07)*. (109). Stockholm.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology*. Chicago: University of Chicago Press.

Stolpe, K., Björklund, L., Lundström, M., & Åström, M. (in preparation). Different profiles characterizing assessment of student thesis in teacher education.

Suto, W. M. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. *British Educational Research Journal, 34*(2), 213-233.

Thurstone, L. (1927). A law of Comparative judgement. *Psychological Review, 34*, 273-286.

Wertsch, J. V. (1998). *Mind as action*. New York: Oxford University Press.

Woolf, H. (2004). Assessment criteria: reflections on current practices. *Assessment & Evaluation in Higher Education, 29*(4), 479-493. doi: DOI: 10.1080/0260293030001689046