Assessment in higher education: exploring the gap between standards and idiosyncracies

Hans Ekbrand Andreas Gunnarsson Maria Jansson
Jan Carle

Department of Sociology and Work science

University of Gothenburg

Over the last 30 years or so, organisations have generally tried to be more transparent and auditable and restrict the trust put on professional discretion and tacit knowledge of their employees, but what room for professional discretion do examiners in higher education still have when they examine student's theses? Using multilevel regression techniques, we use rater effects as an indicator of room for professional discretion and - in particular - the freedom to use idiosyncratic sets of good-making characteristics when evaluating student's theses. The main result is that the between-examiner differences are substantial and much greater than between-department differences and between-faculty differences. Thus we conclude that the departments do not seem to put any significant pressure on the examiner to adhere to a specific idea of what a good thesis is like, nor do the faculties, and that the individual examiners still have a major room for professional discretion in this area.

Introduction

When examiners evaluate examination papers, the individual student's right to equal treatment requires that the evaluation should, in principle, have the same outcome regardless of which examiner was assigned the task of evaluation. In other words, examiners should, in principle, agree on the merits and shortcomings of a particular examination paper.

However, there is a body of research that have shown that examiners are not interchangeable, since there are substantial *rater effects* (Eckes 2008; Wolfe & McVay (2012). Often referenced work in this field are Lumley & McNamara (1995), Weigle (1999) and Barrett (2001). Rater effects have been classified into types, and many studies have tried to explain differences between examiners using properties of the examinators, eg. experience, as predictors (Wolfe & McVay, p 32). In this article we will frame the problem of differences between examiners in a way different from most previous research, that is in the context of the limits

of professional discretion in the current state of academia. We use differences between examiners as a trace of the room for professional discretion of the examiners. The basic idea is that rater effects are indicators of room for professional discretion so that the more rater effects, the larger room for processional discretion.

We will investigate this room for professional discretion in a certain context, ie the context of an interdisciplinary university final theses course within the teachers education programme at the university of Gothenburg. This is a particularily interesting case, since it involves both the individual differences between examiners from the same department and between department differences (and between faculty differences). In a way similiar to the relation between rater effects and professional discretion, we think that systematic between department differences and in particular systematic between faculty differences indicates a sort of independence for the departments and faculties to impose their own versions of scientific quality. Empirical evidence of such differences would be of interest not only to those who try to ensure student's right to equal treatment, but also to researchers in science and technology studies. A careful analysis of the data can reveal both the individual room for professional discretion and how the conceptualisation of what makes a thesis good differs between departments from different kinds of science: the humanities, the social sciences and the sciences.

Purpose

This article will investigate the extent to which the management of a teachers education programme has been able to level disciplinary and institutionally founded differences in evaluation of theses. In addition, we will analyse the how exminers response when faced with organisational pressures to display transparency and accountability, in practice by being instructed to use a campus-common analytic rubric. Particulary, we will investigate the extent of individual idiosyncrasies *compared to* the organisationally founded differnces.

Research questions

Are there any substantial differences between different examiners working within the humanities, the social sciences and the natural sciences, in regard to what criteria they tend to use when evaluating master's theses from teachers education programme? In comparison to between-faculties differences, and between-department differences, what is the level of the between-individual differences in regard to the criteria used in evaluation masters theses?

Theory

Discretion

The core of the concept "profession" includes the authority to make decisions, captured by the concept of discretion. Discretion can be loosely defined as decision making in the absence of surveillance and transparency. Traditionally, university teachers, and examiners, have had substantial discretion, not least when evaluating theses. In the past decades, organisational trends involving codifying standards as a mean of increasing transparency and becoming more 'auditable' have changed matters in most public organisations (Power 1997).

The idea of a university teacher giving only a mark without any reasoning for that decision fits badly into an environment where transparency is celebrated and organisations work on showing themselves auditable. The discretion of the examiner has thus been reduced since organisational demands on transparency and audits require that decisions such as what mark a specific master thesis is given should be a) framed, ie made with a reference to an officially sanctioned list of criteria b) defended by some sort of argumentation.

In a sense, the trust earlier awarded to universities (and other public sector organisations) and their employees to make decisions without having to provide both a set of explicit criteria for the type of decision in question and a description of how the circumstances at hand was evaluated in reference to the criteria, has been redrawn. This trust-redrawal is a general phenomenon manifested in increased demands on *formalisation, transparency* and perhaps most importantly in the education system, *accountability and evaluation* (O'Hara et al 2007, Power 1996). Power (1996) described this change with the following words: "whatever term one prefers, there can be little doubt that something systematic has occurred since 1971. In every area of social and economic life, there is more formalised checking, assessment, scrutiny, verification and evaluation.". Wills and Sandholtz (2009) use the concept "constrained professionalism", which, to some extent, seem to capture the situation of reduced trust that teachers now operate under: "Constrained professionalism represents a new situation in which teachers retain autonomy in classroom practices, but their decisions are significantly circumscribed by contextual pressures and time demands that devalue their professional experience, judgment, and expertise."

Tacit knowledge

Applying professional knowledge on a thesis that is to be evaluated, does not necessarily lend it self well to a general, explicable method. On the contrary, it might well be a much easier task for an examiner to give a well-founded grade of a thesis, than to explain on what grounds the judgment was made. In much the same way, it is more likely that two raters would agree on a grade for a thesis, than that they would provide the same arguments for giving that grade. When giving a grade is no longer consider sufficient, and the raters must also supply a argumentation for the grade, new dimensions of the raters tacit knowledge and personal idiosyncracies will be brought into the fore. On the other hand, the intructions that all the raters in specific course operate under will tend to streamline the evaluations. We understand the position from which the rater is acting as characterised by the pressure to provide an auditable product that in form follows the official requirements - in particular by refering to (all) the criteria, and the pressure to provide a thoughtful and specific commentary on the thesis under scrutiny, and, finally, the requirement to align the overall evaluation with the evaluation of each and every criterion.

Between-rater differences have multiple sources, different ways of managing these pressures as well as different conceptions of what must be fulfilled for a criteria to be satisified. However, in this article we will focus on measuring the extent of between-rater differences, we will not, generally, try to explain them.

Good making characteristics

In an analysis of the usage of the word "good", R. M. Hare developed a logic of value judgements and in particular the concept "good-making characteristics" (Hare, 1952). A good-

making characteristic is a property that makes objects of a certain class good, all else equal. For example good-making characteristics of strawberries could include: being sweet, being red, being big. There are, of course, no general consensus on what properties that are good-making for different classes of objects, not even for strawberries, nor so for student's theses. It could be useful to tentatively think of each examiner having his or her own mental list of good-making characteristics for student's theses, even if that is not really the case. Some examiners might put "clear and creative analysis", and "a comprehensive synthesis of previous research in the field" and a few other elements in his or her list, while other examiners might put in "formulating a relevant and well defined research question" along with "unbiased". There are no such lists, of course, because the examiners use tacit knowledge when they evaluate the theses, and it is, so to speak, the essence of tacit knowledge that it cannot fully be made explicit or written out. Still, the idea of such lists can help us to understand some of the sources for rater effects. Simply put, when examiners apply different sets of good-making characterics in their evaluations of theses, we will find rater effects. In other words, from the perspective of the examiner we can say that being able to apply your own individual set of good-making characterics in the judgement of student's theses indicates that you have a room for professional discretion.

The examiners are instructed to relate to a campus-common analytic rubric when they elaborate their judgement of the theses. This analytic rubric can be seen as an offically approved list of the good-making characteristics of student theses. The function of the rubric, or at least the organisations goal of providing it, is that it provides transparency for students and other interested parties on which criterions that are used in the evaluation, and that it would affect the examiners so that they become more similar in regards to what good-making characteristics of student's theses they acknowledge. Our study can not tell what difference the analytic rubric have done on the examiners ideas of good-making characteristics of student theses, since the rubric has been in use during the whole period under study, but we can pinpoint the differences between examiners that this rubric has not been able to level.

Marking methods

There are two major marking methods, and which one of these the examiners have to use will affect the room for professional discretion: *holistic* marking means that the examiner gives a single measure for the grade, while *analytic* marking implies that the examiner has to provide a grade for each element of the analytic rubric. Holistic marking gives larger room for professional discretion compared to analytical marking - since analytical marking requires the examiner to evaluate every element in the analytical rubric, and leaves no room for other considerations to weigh in - since the final outcome measure is a sum of the listed grades only. This way the examiner are hindered not only from using other good-making characteristics than the officially approved one, but also to give them different weights. Of course, a clever examiner could by pass this curbs by simple reducing the scale of variation of the awarded grades on the characteristics that s/he does not find very important, and save the really high and low grades for the more important ones. Still, analytic marking does tend to reduce the freedom for the examiner, and reduce inter-rater differences.

> [...] analytic marking is likely to improve inter- and intra-rater reliabilities. With holistic scoring, although particular traits are specified, raters may include traits not listed in the marking scheme and/or 'use personal judgment to determine how important a specific trait is to the overall score' (74). This can lead raters to

> move away from the criteria originally designed to define what is being assessed and, consequently, reduce score consistency within and across raters. (Barkaoui 2011, quoting Goulden 1994).

In the case at hand, the examiners was instructed to use holistic discretion and give only one grade. The grade must however also be accompanied by a motivation which - according to the instructions - should include an evaluation of the elements of the analytical rubric.

Previous research

In an overview of research on the effects of the use of rubrics in assessment, Rezaei & Lovorn (2010) found that previous studies consistently come to the same conclusion - the use of rubrics makes assessment more reliable, i.e. it reduces the rater effects. At the same time, rubric use has been critised on the grounds that it leads to too narow concepts of quality (Wilson 2007). Rezeai & Lovron (2010) did a series of experiments to estimate the effects of rubric use on the validity of the assessments. While there were some effects, the expected decrease in variability of grades was not one of them. The raters did not grade the elements in the rubric in the way the experiment designers had expected - e.g. rater could give points for "citations and references" even when there were no citations or references in the rated essay (ibid. 27). One of the conclusions was that proper training in the use of the rubric is needed in order to get the expected benefits in terms of reduced rater effects, a conclusion also supported by a study reported in Boulet et al. (2004), and in a review of the litterature (Malini Reddy & Andrade 2010).

In a survey study where raters would report the importance the gave to a list of criteria, Thomas Eckes found empirical support for sex different rater types, where each type represented a certain configuration of what importance the raters would give to the criteria (Eckes 2007). While the specific rater types found in an empirical study to some degree depends on the domain of the assessment practice - which in Eckes study was writing performance in a second langauge - some differences between the types might represent differences between raters more generally. For example one of the rater types was predominantly occupied with issues of form (syntax), and Eckes showed that age of the rater was correlated to this type, with younger raters more likely to put weight to syntax issues. On a more general level, Eckes results contributes the litterature that have shown that there are substantial rater effects.

(To be expanded).

Data

This study is based on written evaluations that are provided by raters as part of the examining process of theses from a teachers education programme at university of Gothenburg in Sweden. A draft version of these documents are handed over to the students when they have defended their thesis, and the final and official version is later archived together with the final version of the thesis. While these documented evaluations vary in length and form, they manifest the official judgement of the merits and shortcommings of the theses. As such, they are cultural artefacts that can tell us things about their authors. While these texts could also be used as source of knowledge about the theses, in this article we concentrate solely on them as a source of knowledge about their authors, the raters.

In addition we also have access to data about the raters, e.g. in what department they work, gender, age, academic title, date of PhD-title, number of examinations done prior to the current examination.

Lastly, the grade of the thesis is also part of our data matrix.

Methodology

Operationalisations

Compliance is measured as how many of the criteria are refered to in the written evaluation. The more of these criteria refered to, the more compliant the rater. When we say that a particular examiner tends to put greater weight in a certain criterion than another examiner, we refer to the number of words the former uses when commenting the criterion in question. The rationale for this operationalisation is the idea that the evaluation can be seen as a restricted text in which different remarks about the thesis compete for their space. While there is no formal restriction on the length of the evaluation, the rater has limited time avaible to use for writing it, so he or she will prioritize his or her most important remarks during writing of the evaluation.

We have coded each evaluation by applying different sets of codes to each sentence of text. The first set of codes concerns the criteria that is relevant in the text, and we call these codes *criteria-codes*. The rater does not need to use a particular term (e.g. the name of the criterion in the list of criteria), since we code according to how we interpret the particular sentence. Several codes may well be applied to a particular sentence, since it is possible make several remarks in one sentence. Our code scheme when it comes to the critieria is essentially a reduced version of the official list of criteria for the course, where some of the criteria have been aggregated into one code. The codes we use are: (1) language and disposition; (2) aim and research questions; (3) theory and previous research; (4) method and research ethics; (5) analysis, conclusions and discussion; (6) the overall impression and the integration of the parts into a coherent work, (7) implications for practice.

The second set of code concerns the discursive function of the sentence, where we use the following categories: (a) evaluative, (b) descriptive, (c) questioning, (d) explaining and (f) imperative. This set of codes we call *mode-codes*.

Analytic strategy

The first research question will be analysed using two matrices. The first matrix is a tabulation of criteria-codes and faculties, where the text of each rater is aggregated under the faculty in which he or she works. The content of each cell of this matrix represents the proportion of text written by the staff of the faculty that is coded with this particular criteria-code. By means of an anova test we then determine if there are substantial between-faculty differences in what criteria are given higher priority. The same technique will be used to analyse the mode-codes.

The second research question will be tackled with a multi-level regression framework, in order to partition the unexplained variation between the faculty level and the individual level.

Findings

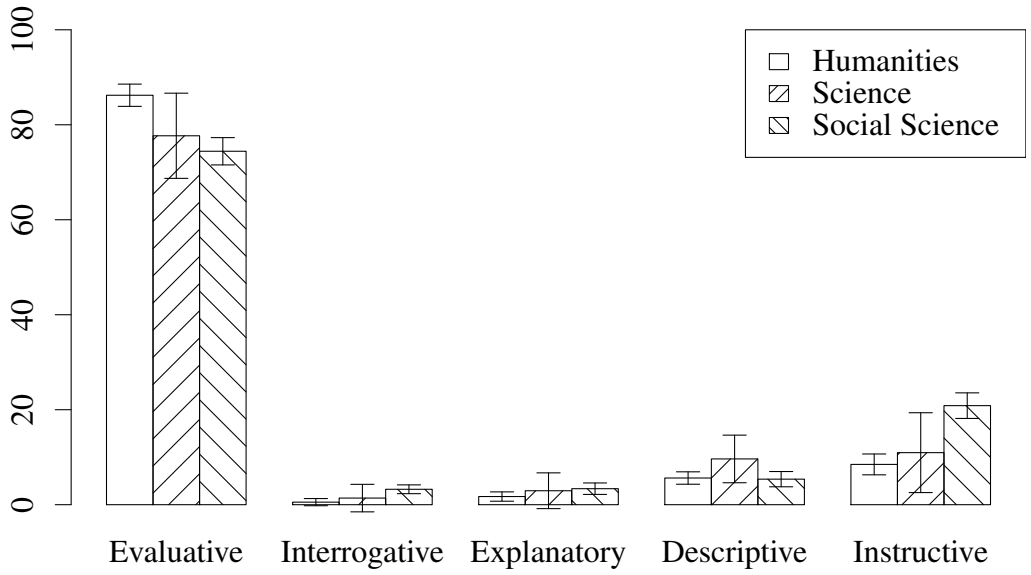
Differences between faculties

Let us first have a look on the use of text mode-codes. Table 1 shows, in per cent, the distribution of text mode codes in the evaluations given by teachers from the humanities, the social sciences and the sciences respectively. The figures are percentages and the base for these calculations is the number of words written by examiners from each faculty. E.g. 84.10 per cent of all words written be examiners from the humanities appear in sentences that were coded as "Evaluative" while only 0.50 per cent of the words written by these examiners appeared in sentences codes as "Interrogative".

Table~1: Faculty differences in the use of text mode codes.

	Evaluative	Interrogative	Explanatory	Descriptive	Instructive
Humanities	84.10	0.50	1.70	5.50	8.30
Science	75.70	1.40	2.90	9.40	10.70
Social Science	69.40	3.00	3.10	5.00	19.40

The confidence intervals around the means presented in table 1 is shown in graph 1.



The social sciences stands out in that the examiners from these disciplines tend to ask questions to the students and provide instructions on how the students should improve the work examined. Examiners from the humanities and science prioritize giving jugdements more often than examiners from the social sciences. There are rather small differences between the humanities and science, mainly that explaining is relatively common in science but relatively uncommon in the humanities.

Next, we turn to criteria-codes. The criteria-codes are divided into sub-categories according to the kind of judgement they represent: positive, negative, neutral or missing. When an examiner use a negative criteria-code that does not necessary mean that the work fails. [Exempel på hur det kan låta här, som visar att det inte är så allvarligt att få kritik] Table 2 shows the use of negative criteria-codes. The base of the per cent calculations is the total word count of all negative criteria-codes.

Table~2: Faculty differences in the use of negative criteria codes

	Method	Aim	Analysis	Prev.Research	Language	Relevance	Total
Humanities	13.94	11.59	24.19	20.51	27.00	2.76	100.00
Science	18.26	8.74	33.15	10.66	26.11	3.07	100.00
Social Science	19.96	15.26	19.83	16.53	27.17	1.25	100.00

Examiners in Science more often than others point out problems with analysis and lack of relevance (for the teaching profession), while examiners from the humanities focus relatively often on problems concering the relation to previous research. Remarks about language problems are common to examiners from all faculties. Lastly, examiners from the social sciences is more likely to discuss problems with methodology and the aim.

Between-rater differences – manifestations of discretion

If there were no between-rater differences, nor any between departments, or between-faculties differences in regard to the use of text-modes or criteria codes, we would not be able to predict either of them using information about the rater and department and the faculty of the rater. All differences in the evaluations could then be attributed to the properties of the thesis per se. However, that is not the case. We can indeed predict, to some extent, the use of certain text-modes as well as the use of certain criteria-codes using information about which examiner wrote it, or the department or faculty of the examiner. Actually, the proportion of variation that can be attributed to the individuals represents the discretion that we are interested to estimate. And the proportion of variation that can be attributed to the department and faculty levels, i.e the between-departments and between-faculty differences, represents the idiosyncrasies characteristic for the different faculties that the levelling efforts of the course administrators has not yet been able to erase.

Table 3 shows, in per cent, the variation in the use of text-modes that can be attributed to the examiner, the faculty and the thesis per se (“unexplained”).

Table~3: Partitioning of the sources of variance for mode of text

	Examiner	Department	Faculty	Unexplained	Total
Evaluative	27.42	1.48	4.77	66.32	99.99
Interrogative	11.33	0.00	1.87	86.79	99.99
Explanatory	9.96	0.67	0.00	89.36	99.99
Descriptive	16.11	3.00	0.00	80.89	100.00
Instructive	35.42	2.46	3.03	59.09	100.00

While there are substantial differences between the values within the same column, we are interested in the differences between columns. Most of the variation is unexplained, or in other words, can be attributed to the work *per se*. On average, about 15 per cent of the variation is due to individual idiosyncrasies. In comparision, the variation attributable to the between-faculties differences is very small, only a few percentages.

The analysis of negative criteria codes are shown in table 4. Looking at the column "examiner" we see that remarks about bad language is the area in which the individual discretion is highest – almost 25 percent of the variation in language remarks are due to individual idiosyncracies among the examiners. On the other end at the discretion-continuum we find relevance (in this case, relevance for the teaching profession) were as much as 95 per cent of the variation is unexplained (or due to properties of the work).

Table~4: Partitioning of the sources of variance for criteria use in negative remarks.

	Examiner	Department	Faculty	Unexplained	Total
Method	6.66	0.84	4.48	88.01	99.99
Aim	3.47	0.00	2.90	93.63	100.00
Analysis	14.98	0.00	0.94	84.09	100.01
Prev Research	7.63	1.48	0.77	90.12	100.00
Language	24.61	1.41	1.08	72.90	100.00
Relevance	3.89	1.18	0.00	94.93	100.00

When we say that only 5 per cent of this variation is due to the examiners, that means that negative remarks about lack of relevance for the teaching profession seems to be almost randomly spread over evaluations – that an examiner has made such a remark once does not increase the chance of him or her doing it again. In contrast, negative remarks about language seem to make up relatively high proportions of the evaluations for some examineres, while other examiners have relatively few such remarks. This points to the hypothesis that there might be a type of examiner that – in comparision to his or her collegues – focus markedly more on languaguge than on other criteria-codes.

The between-faculty differences shown earlier almost disappear when controlling for individual syncrasies, for all criteria-codes but methodology and aim. That means that within each faculty, there are great between-rater differences, except when it comes to remarks about methodology and aim. [titta närmare på vilken fakultet som skiljer ut sig, det bör vara samhällsvetenskapliga fakulteten]

Since the column "Unexplained" shows the relative importance of the thesis *per se* for the propabilities that a negative remark is awarded, we can conclude that a negative remark about "langauge" to a larger extent than a negative remark about (lack of) "previous research" can be attributed to other things than the thesis *per se*. Or, which is just another aspect of the same phenomenon, we can say that the criterium for "language" is the criterium which seem to be least consensus about among the examiners.

In table 5, the use of criteria codes in positive remarks is analysed, much in the same way as table 4 did for the negative remarks.

Looking at the "Unexplained" column, we can conclude that a positive remark about "analysis" is substantially more meaningful than a positive remark about "relevance". (Perhaps this says something important about notion of "science" that the examiners apply in their assessments?)

Table~5: Partitioning of the sources of variance for criteria use in positive remarks.

	Examiner	Department	Faculty	Unexplained	Total
Method	26.12	3.13	1.03	69.72	100.00
Aim	23.08	0.00	0.00	76.92	100.00
Analysis	15.06	0.25	0.80	83.89	100.00
Prev Research	22.26	2.35	0.00	75.39	100.00
Language	24.62	2.29	0.00	73.09	100.00
Relevance	34.23	0.00	0.58	65.19	100.00

When we compare table 4 and table 5, we see that the proportion of unexplained variation is higher for the negative remarks than for the positive remarks. (I think this relates to that assessment is about being "good enough").

Discussion

There are substantial between-faculty differences when it comes to what the examiners gives the students in terms of imperatives, questions and explanations. Such elements makes evaluations more formative than the evaluative and descriptive text modes. Examiners from the social sciences tend to direct their effort in this direction, more often than examiners from the humanities or science who instead focus more on evaluative elements.

In regard to criteria-codes, examinators from science stand out as more interested in questions about analysis, compared to examiners from the social sciences or the humanities. The preoccupation with methodology and aim is characteristic for the social sciences. A rather unexpected result is that examiners from the humanities more often then others complain about how students relate to previous research.

The individual examiners have a substantial discretion when assessing theses. The departments do not seem to put any significant pressure on the examiner to adhere to a specific way of doing assessments, nor do the faculties. However, this is not to say that the examiners are not governed at all - they still work under a common framework given by the management of a teachers education programme and the effect of that common framework is not tested in this study. But the subcultures of the departments or faculties do not leave much of a trace in the assessment practice. The differences between examiners in the same department overshadow the differences between departments.

References

Barkaoui, Khaled (2011) Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, 18:3, 279-293.

Boulet, J.R, Rebbecchi, T.A., Denton, E.C., Mckinley, D., & Whelan, G.P. (2004) Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education* 9: 47–60.

Eckes, Thomas (2008) Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155–185.

Goulden, N.R. 1994. Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education* 27, no. 2: 73–82.

Malini Reddy, Y., Andrade, Heide 2010 A review of rubric use in higher education *Assessment & Evaluation in Higher Education*, 35, no. 4: 435-448.

Rezaei, Ali Reza & Lovorn, Michael (2010) Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15, 18–39.

Wolfe & McVay (2012) Applications of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practices* 31:4, 31-37.

Wills, John & Sandholtz, Judith Haymore (2009) Constrained Professionalism: Dilemmas of Teaching in the Face of Test-Based Accountability. *Teachers College Record* 111:4, 2009, p. 1065-1114.

Wilson, M. (2007) Why I won't be using rubrics to respond to students' writing. *English Journal*, 96, 62–66.