UNIPED

UNIVERSITETSFORLAGET

RESEARCH PUBLICATION

ÅRGANG 41, NR. 3-2018, S. 319–330
ISSN ONLINE: 1893-8981
DOI: 10.18261/ISSN.1893-8981-2018-03-11

Grade Correspondence between Internal and External Examiners of Occupational Therapy Students' Bachelor Theses

Karaktersamsvar mellom interne og eksterne sensorer på ergoterapistudenters bacheloroppgaver

Tore Bonsaksen
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway
Faculty of Health Studies
VID Specialized University, Sandnes, Norway
tore.bonsaksen@oslomet.no

Mikkel M. Thørrisen
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway
mikkel-magnus.thorrisen@oslomet.no

Unni Sveen
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway
Department of Physical Medicine and Rehabilitation
Oslo University Hospital, Oslo, Norway
unni.sveen@oslomet.no

Ingvild Kjeken
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences, OsloMet – Oslo Metropolitan University, Oslo, Norway
National Advisory Unit on Rehabilitation in Rheumatology, Department of Rheumatology, Diakonhjemmet Hospital, Oslo, Norway
ingvild.kjeken@oslomet.no

Randi W. Aas
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences,
OsloMet – Oslo Metropolitan University, Oslo, Norway
Department of Health Studies, Faculty of Social Sciences
University of Stavanger, Stavanger, Norway
randi.aas@oslomet.no

Anne Lund
Department of Occupational Therapy, Prosthetics and Orthotics, Faculty of Health Sciences,
OsloMet – Oslo Metropolitan University, Oslo, Norway
anne.lund@oslomet.no

ABSTRACT
Students' grades are increasingly important in defining their future employability, and therefore, securing a fair assessment of students' theses is important. This study aimed to assess the level of grade correspondence between internal and external examiners of occupational therapy students' bachelor theses, and to evaluate the overall level of grades initially set by examiners in the two groups. The grades initially suggested for 67 bachelor theses were analyzed. Absolute agreement between internal and external examiners was estimated as the proportion of theses on which the examiners suggested identical grades (percentage agreement), and consistency in agreement was estimated by the intraclass correlation coefficient (ICC). There was absolute agreement between the internal and the external examiners in 33 of the 67 cases (49.3 %), and also consistency in agreement was high (ICC = 0.81, 95 % CI [0.68, 0.88], $p < 0.001$). The results from this study demonstrate a high level of agreement between internal and external examiners of occupational therapy students' bachelor theses. However, internal examiners as a group are more prone to give high grades compared to external examiners, and this may support the continued use of two examiners to ensure quality in grading.

Keywords
assessment, grading, interrater reliability, undergraduate students

SAMMENDRAG
Karakterene som studenter får er vesentlige for fremtidige ansettelser. Derfor er det viktig at vurderingen av studentenes bacheloroppgave i siste studieår blir gitt på en rettferdig måte. Hensikten med denne studien var å vurdere nivået av karaktersamsvar mellom interne og eksterne sensorer på ergoterapistudenters bacheloroppgaver, og å vurdere det overordnede karakternivået, slik dette var foreslått av intern og ekstern sensor før samsensur. De foreslåtte karakterene for 67 bacheloroppgaver ble analysert. Absolutt karaktersamsvar mellom intern og ekstern sensor ble estimert som andelen av oppgavene hvor de to sensorene foreslo eksakt lik karakter (prosent enighet), mens konsistens i karaktersamsvar ble estimert ved intraklassekorrelasjon (ICC). Det var absolutt enighet mellom de to sensorene i vurderingen av 33 av de 67 oppgavene (49.3 %), og konsistensen i karaktersamsvar

var også høy (ICC = 0.81, 95 % CI [0.68, 0.88], p < 0.001). Resultatene fra studien viser et høyt nivå av samsvar mellom interne og eksterne sensorers vurdering av bacheloroppgaver i ergoterapi. Interne sensorer er mer tilbøyelige til å gi bedre karakterer enn eksterne sensorer. Dette støtter fortsatt bruk av to sensorer for å sikre kvalitet i karaktersettingen, og fornyet innsats for å bedre samsvaret mellom sensorene.

Nøkkelord
bacheloroppgave, ergoterapi, høyere utdanning, interrater reliabilitet, karaktersetting, vurdering

INTRODUCTION

Assessment has been emphasized as an important element in higher education, benefiting students as well as society and its institutions of higher education. Two main forms of assessment have been identified. On one hand, formative assessment denotes "assessment for learning," i.e. assessment conducted with the purpose of supporting the students' learning (Wiliam, 2011). This form of assessment is typically used as an integral part of the teaching in a course. On the other hand, the traditional summative assessment denotes assessing what the student has learned, and is therefore typically used after the completion of a course. Summative assessment of students' performance – often expressed by a grade – is a necessary aspect of policy makers' need to control and evaluate learning institutions and society's use of resources (Scriven, 1991). For the students, grading has an impact on their learning by directing attention to what is important, as well as by acting as incentives for studying (Boud & Falchikov, 2007). Grades are increasingly important in defining students' future employability, and assessment of final-year thesis or dissertation is of particular importance as it tends to contribute substantially to degree classification (Saunders & Davies, 1998).

Grading and assessment raise several issues. Assessment criteria should be clear, understandable and closely associated with predefined learning outcomes (Baume, Yorke, & Coffey, 2004; Saunders & Davies, 1998). The grading processes should be valid, reliable, feasible and fair (Calvert & Casey, 2004; Hand & Clewes, 2000; Lomas & Nicholls, 2005; Newstead, 2002). "Assessment cultures" may, however, vary considerably between fields, institutions and departments (Wolf, 2004), and individual examiners may differ in how they utilize and apply assessment criteria. Whereas some rely on overall (holistic) evaluations of the student work, others may obey more rigidly to predefined assessment critera (Grainger, Purnell, & Zipf, 2008).

According to Sadler (1989), one may differentiate between four specific criteria relevant to academic work: (a) relevance to the task, (b) logical development and validity of arguments, (c) clarity of expression (organization), and (d) technical aspects (presentation). Overall evaluations of written academic work are thus complicated by individual differences between examiners regarding how to weight different aspects of the product. Grainger and colleagues (2008) conducted a study among a group of examiners in order to explore what common criteria examiners apply when assigning grades and whether they have a similar interpretation of what constitutes "quality". Even though they found that the

examiners did have a quite common set of assessment criteria (emphasis on content knowledge and technical aspects), the examiners displayed dissimilar interpretations of "quality" related to these criteria. Consequently, Grainger and colleagues (2008) emphasized the need for increased consensual agreement on descriptors of standards among examiners.

Different approaches to consensus regarding grading coexist. Johnston (2004) proposed that one may distinguish between positivistic and interpretative approaches. The goal of a positivistic assessment entails identifying a "true" grade for the student, based on the premise that there indeed exists an objectively correct grade that reflects the student's level of knowledge, competence and performance. An interpretative approach, on the other hand, would reject this notion, viewing grades as a result of systematic consideration of the students' product in light of relevant assessment criteria, but without proposing that the grades reflect an objective reality.

Assessing complex assignments, such as theses wherein students quite freely choose their own topics, is particularly difficult as the grading process requires the examiner to make qualitative judgements in lack of detailed assessment criteria (Rasch & Eriksen, 2008; Sadler, 2013). In such situations, a written guide for examiners may be helpful, and many higher education institutions do in fact distribute such guidelines to both internal and external examiners. However, using two examiners may be a particularly serviceable measure in order to maximize assessment validity and reliability. Hence, the question of correspondence between examiners arises. Although evidence is somewhat mixed, research has generally suggested challenges related to establishing inter-examiner agreement on student work in higher education, particularly with regard to essays and other forms of complex academic student work, such as bachelor theses (Asmyhr, 2011; Bettany-Saltikov, Kilinc, & Stow, 2009; Bjølseth, Havnes, & Lauvås, 2011; Larsen, Johnsen, & Pallesen, 2006; Lauvås & Jakobsen, 2002; Rasch & Eriksen, 2008). When studying correspondence between internal and external examiners on bachelor students' essays at a Norwegian university college, Asmyhr (2011) uncovered differences between examiners that surpassed two levels on a six-level grading scale. Although available, explicit assessment criteria were utilized by the examiners to a small extent. Rather, they tended to apply a interpretative and holistic approach to assessment. Rasch and Eriksen (2008) explored the extent to which grade suggestions from a first examiner influence the assessment made by a second examiner. They found that an erronous assessment made by the first examiner tended to influence and bias the assessment made by the second examiner, rather than being corrected by the latter.

On the other hand, Bettany-Saltikov and colleagues (2009) found good inter-examiner reliability when studying correspondence between examiners on master level projects at a university in the United Kingdom. A set of examiners was provided with projects that had already been graded, and variations in grades from the original assessments did not exceed 6 percent on average. Similarly, Larsen and colleagues (2006) uncovered a high degree of consensus between examiners on assessments of undergraduate psychology students' essays at a Norwegian university, with an average intraclass correlation of 0.83 across 13 exam commissions, each consisting of two examiners. In summary, the evidence related to grade correspondence between examiners is mixed,

and the current study appears to be the first to examine this issue in the context of occupational therapy education.

STUDY AIMS AND RESEARCH QUESTIONS

The study's primary aim was to assess the level of initial grade correspondence between internal and external examiners. In cases where there was disagreement between the two examiners, we examined how the final (given) grade related to the two initial grades suggested by the internal and external examiner. However, we also compared the two groups of examiners with regard to their initial grading. Building on the stated aims, the research questions were:

- What is the level of grade correspondence between (a) pairs of internal and external raters, and between (b) the groups of internal and external raters?
- In cases of disagreement, is the final grade assigned to the student closer to the internal rater or the external rater's initial grade?

METHODS

Context and design

This study was conducted at the occupational therapy education program at Oslo and Akershus University College of Applied Sciences in Oslo, Norway. Approximately 250 students are enrolled in the program, and about 70 students graduate on an annual basis (Bonsaksen, Kvarsnes, & Dahl, 2016). The education program is an undergraduate program with a duration of three years encompassing 12 study modules (Oslo and Akershus University College of Applied Sciences, 2011).

In the last module of the program, pairs of students write their bachelor thesis in collaboration. With special reasons, students may be allowed to write alone, subject to administrative approval of the student's application to do so. The thesis is a scholarly work where the problem to be addressed is related to current research and/or development in the field of occupational therapy and/or occupational science. The time assigned for conducting the project and writing the thesis is approximately 10 weeks. The learning outcomes include having knowledge about science and scientific methods of inquiry, demonstrating the ability to use that knowledge in a supervised research process, and demonstrating the ability to report and discuss the findings of the conducted inquiry (Oslo and Akershus University College of Applied Sciences, 2011).

Ultimately, the theses are graded according to the general grading system in higher education in Norway (The Norwegian Association of Higher Education Institutions, 2011). Thus, the grades reflect a student performance described as excellent (A), very good (B), good (C), satisfactory (D), sufficient (E), and fail (F). Table 1 displays the grade descriptions in more detail. Two examiners, one teacher at the education program and one external to the program, grade each thesis by coming to agreement. The internal examiners also function as the students' supervisors.

Table 1. The general qualitative descriptions of grades in Norwegian higher education

Symbol	Description	Qualitative description of valuation criteria
A	Excellent	An excellent performance, clearly outstanding. The candidate demonstrates excellent judgement and a high degree of independent thinking.
B	Very good	A very good performance. The candidate demonstrates sound judgement and a very good degree of independent thinking.
C	Good	A good performance in most areas. The candidate demonstrates a reasonable degree of judgement and independent thinking in the most important areas.
D	Satisfactory	A satisfactory performance, but with significant shortcomings. The candidate demonstrates a limited degree of judgement and independent thinking.
E	Sufficient	A performance that meets the minimum criteria, but no more. The candidate demonstrates a very limited degree of judgement and independent thinking.
F	Fail	A performance that does not meet the minimum academic criteria. The candidate demonstrates an absence of both judgement and independent thinking.

The study was designed as a retrospective inter-rater reliability study. It was conducted in Oslo, Norway, during 2015 and 2016. Each of the years, pairs consisting of seven internal and seven external examiners rated occupational therapy bachelor theses. Over the two years period, the theses of approximately 130 students were assessed.

All of the assigned examiners were informed about the study and volunteered to participate. Each pair of examiners graded four or five theses. All examiners satisfied the minimum education criteria of having at least a master's degree. Almost all the internal examiners had a Ph.D. degree, had approved competence equal to this level, or was a Ph.D. student.

Data production and collection

A total of 67 bachelor theses, 35 submitted in 2015 and 32 in 2016, constituted the data material in this study. Participation in the study meant that all examiners would provide a preliminary grading, solely based on their own judgment, on each assigned thesis. Afterwards, the two examiners would meet (in person or by telephone), discuss their initial ratings, and reach an agreement on grading. The internal examiner's, the external examiner's, and the agreed-upon grades were all registered on a special form developed for this study.

Additionally, we registered the cases where students required a justification for the grade they had been given. We also registered the cases where the students demanded a second evaluation of their thesis, and the result of the second evaluation. After collecting all the completed forms, the information was transmitted to IBM SPSS and checking for mistakes was performed. All data were collected anonymously. Approval from the Norwegian Data Protection Official was therefore not required.

Analysis

For the statistical analyses, the grades denoted with letters were transformed as follows: A (best grade)=6, B=5, C=4, D=3; E=2, and F (lowest grade; fail)=1. In order to estimate the agreement between the internal and external examiners, two methods were applied. First, we calculated the proportion of the theses on which the internal and the external examiners had identical preliminary grades (percentage agreement; PA). Next, intraclass correlation coefficients (ICC) were produced (Shrout & Fleiss, 1979; Streiner & Norman, 2008). We used a mixed-effect model treating theses as fixed factors and examiners as random factors. The ICCs can be calculated in two ways. One can produce a measure based on the two raters' absolute agreement; i.e. the extent to which the raters suggested the exact same grade. Alternatively, one can produce a measure of the two raters' consistency in agreement; i.e., the extent to which a high grade given by one rater corresponds with a similar high grade (but not necessarily the exact same grade) by the second rater. Considering that absolute agreement on grades appears to be an unreasonably high standard, we were interested in the examiners' consistency in agreement, not only their absolute agreement; thus, we used the consistency type. The ICC is interpreted similarly to well-known measures of reliability, like the Cronbach's α. Essentially, $\alpha \geq 0.70$ is generally considered acceptable, whereas $\alpha \geq 0.80$ is preferred (Streiner & Norman, 2008). A 95 % confidence interval was constructed around the ICC average measure.

Descriptive analyses with frequency tables were produced in order to provide an overview of the grade distribution in the two groups of examiners, and the way that grades varied between them. The Mann-Whitney U-test was used to examine the overall grade level as suggested by the internal and external examiners, respectively. Requests for grade justifications and requests for a second evaluation were analyzed descriptively. For the inferential analyses, the level of statistical significance was set at $p < 0.05$.

RESULTS

Examiner agreement

There was agreement between the internal and the external examiners in 33 of the 67 cases, resulting in 49.3 percent absolute agreement. The internal examiners suggested one level higher grades in 23 cases (34.3 %), two levels higher in nine cases (13.4 %), and four levels higher in one case. In only one case did the external examiner suggest a grade one level higher than the internal examiner. When assessing the consistency type agreement, we found a high level of agreement between the internal and the external examiners (ICC = 0.81, 95 % CI [0.68, 0.88], $p < 0.001$).

In the 34 cases of disagreement between examiners, the final (given) grade was closest to the grade initially suggested by external examiner in 16 cases (47.1 %). The two examiners agreed on the average between the two initial grades in 8 cases (23.5 %), whereas the final grade was closest to the grade suggested by the internal examiner in 10 cases (29.4 %).

Differences between internal and external examiners

Table 2 shows the frequency and proportion of grades as initially suggested by the internal and the external examiners, respectively. Overall, the internal examiners suggested grades that were significantly higher than the grades suggested by the external examiners ($p < 0.01$).

Table 2. Frequency and proportion of grades as initially suggested by the internal and external examiners

	Internal examiner		External examiner	
Grade	**n**	**%**	**n**	**%**
A	17	25.4	9	13.4
B	26	38.8	20	29.9
C	18	26.9	23	34.3
D	3	4.5	7	10.4
E	2	3.0	7	10.4
F	1	1.5	1	1.5
Median grade	5		4	

Note. A median numerical grade of 5 corresponds with the actual grade B, whereas the median numerical grade of 4 corresponds with the actual grade C.

Discussion

The main aim of this study was to explore the level of grade correspondence between internal and external examiners of occupational therapy students' bachelor theses. The results demonstrated a high level of agreement, with an ICC of 0.81 ($p < 0.001$) between the examiners. There was an absolute agreement in half of the 67 cases, and only one grade of difference in 23 of the remaining 34 cases.

Looking at the 34 cases in which the examiners differed in their evaluation, there was only one case where the external examiner suggested a higher grade than the internal examiner. Thus, the results revealed a pattern in which internal examiners conistently judge the quality of the bachelor theses as better than the external examiners do. Thus, there may be different "assessment cultures" between internal and external raters (Wolf, 2004). One aspect of such cultural differences may concern expectations towards the students and their performance. The internal examiners have followed the students' development through teaching and supervision over a longer period of time and are more familiar with the curriculum of the occupational therapy education than the external examiners. As a consequence, they may have more modest expectations to what knowledge the students should have gained after three years of education. External examiners, on the other hand, may have a more distant relation to the educational system, and may find it difficult to differentiate between the knowledge one may expect to be gained through the bachelor level education, and knowledge developed as a consequence of clinical experience and later education at a master or Ph.D.-level. They may therefore have higher expectations regarding

quality and thus be stricter in their evaluations. This hypothesis is in line with Sadler's (2005) suggestion that the dominant approach to judging the quality of student work is based on a combination of the marker's personal expectations and the way in which the student has performed in relation to other students, the latter often called norm referencing. Since internal examiners have more knowledge about the general level among students, they may be more prone to norm referencing than external examiners, who to a larger degree may apply criteria-based assessment.

Another explanation for the conistent pattern of higher ratings given by internal examiners relates to their role as supervisors for the students they later grade. Thus, the rating may also include an element of grading one's own degree of success as a supervisor, which may lead to a tendency of overrating the quality of the theses. One way to explore this hypothesis in future studies may be to compare agreement between ratings performed by two external examiners with agreement between ratings performed by an external and an internal examiner who also act as a supervisor.

A third explanation for the higher ratings among the internal examiners may be that the external examiners to a larger degree based their judgement on the predefined assessment critera, thereby applying an analytic marking approach. The internal examiners may use a more interpretative and holistic approach in their judgement, taking into account also their knowledge about the student's learning process and amount of effort given to the work with the thesis. This hypothesis corresponds with the results from a study evaluating the reliability of generic assessment criteria when used by lecturers from different disciplines (Bettany-Saltikov et al., 2009). In that study, there was a general agreement among the participants that marking is affected by the relationship to the student. The reason for this may be that supervisors can often see the effort that the student has put into their thesis, or have a particular interest in the topic area. However, and in contrast to the results from our study, the participants argued that this can work in both a positive and negative manner. Interestingly, Bettany-Saltikov and colleagues noted that in those cases where the second examiner had marked the thesis significanty lower than the supervisor, the participants in the study agreed with the supervisors' marks (Bettany-Saltikov et al., 2009).

Even if the agreement between examiners in general was high in our study, there was a discrepancy of two levels in nine cases and four levels in one case. This suggests that a more detailed set of standards descriptors for each grade may be needed. The existing descriptors (see Table 1) are rather brief and provide limited guidance. According to Wolf (2004) the formulations used to define various grades may not be clear and concise enough, they may even be regarded as "unhelpful formulations". Another aspect of dicrepancy between examiners may be due to the assessors holding different concepts of quality, as suggested by Sadler (2013), as well as differ on how the symbols should be assigned (Sadler, 2013). In the previously mentioned study of the process of decision-making by multiple markers assessing the same students, the results revealed that even if the markers shared a common understanding of quality in the context of the marking criteria and standards, they awarded different levels of achievement in some cases (Grainger et al., 2008). The same results were found in a study by Bjølseth and colleagues (2011), where examiners, after having participated in six workshops with rating exercises and discussions over a one year period, still showed large variability in their assessments, with an absolute agreement of

only 6 % and a discrepancy of two levels or more in 81 % of the cases. It is therefore debatable whether assessment can be totally objective and free of the social and individual conditions in which it is practiced. Thus, some disagreement between examiners is to be expected.

The cases with initial disagreement between the internal and external examiners' grade suggestions are interesting in terms of whose initial opinion would have the most impact on the final given grade. An agreement on the average between the initially suggested grades appears to be the preferred decision in cases where the initial difference was two grade levels. This solution was applied in 24 % of these cases. Among the other cases of initial disagreement, the internal examiner's initial rating was weighted most strongly in 29 % of these cases, whereas the external examiner's initial rating was weighted most strongly in 47 % of these cases. This implies that there is a tendency of leaning towards the external examiner's opinion in cases of disagreement, a practice that also is in line with the expressed assessment tradition at the university. Moreover, if there is a tendency of norm-based opinions among internal examiners, as argued (Sadler, 2013), there is good reason to continue such grading practice in cases of examiner disagreement.

However, the final grade may be much influenced by which of the two examiners present their initial grade first. As Rasch and Eriksen (2008) found, there may be a tendency that the examiner who first presents his assessment exerts more influence on the student's final grade than the second examiner. Taken together, one may suggest that two examiners can assist in upholding the quality of the final grade given to students, provided the examiners find a way to minimize the risk of leaning towards the first-expressed opinion.

Study limitations

One limitation of our study is the lack of demographic information about the examiners concerning gender, age, years of experience or academic position. Studies have demonstrated that professors tend to be stricter than others (Rasch & Eriksen, 2008), and that long experience does not lead to more consistent ratings (Bjølseth et al., 2011). We also analyzed a relatively small amount of bachelor theses, from one field of education, and from only one education institution. These are all aspects that limit our ability to generalize the findings to the higher education field in general.

Conclusion and future studies

The high level of agreement between internal and external examiners of occupational therapy students' bachelor theses is promising, but the higher overall grade level suggested by the internal examiners supports the use of two examiners to ensure quality in the grading of bachelor theses. Alternatively, one could suggest that supervisors should not serve as internal examiners on the theses they had supervised. Future studies should explore factors that may influence grading, and in particular the process by which the internal and external examiners reach their final grade conclusion. It would be interesting to know if the use of written guidelines for examiners would contribute to a higher level of agreement on the grades they suggest. Finally, we suggest that future studies explore the possibility of establishing and sustaining an assessment culture, a culture which may contribute to a higher level of agreement on grades between internal and external examiners. Such an assessment

culture may also extend across higher education institutions, which may be a substantial step forward in terms of securing that grades reflect the same levels of competence regardless of where they have been given.

REFERENCES

Asmyhr, M. (2011). Om vurdering av essaybesvarelser i høyere utdanning – en studie av vurderer-reliabilitet. *Uniped, 34*(4), 17–33.

Baume, D., Yorke, M., & Coffey, M. (2004). What happens when we assess and how can we use our understanding of this to improve assessment? *Assessment & Evaluation in Higher Education, 29*(4), 451–477. DOI: http://dx.doi.org/10.1080/0260293031000168903 7.

Bettany-Saltikov, J., Kilinc, S., & Stow, 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. DOI: http://dx.doi.org/10.1080/02602930802302196.

Bjølseth, G., Havnes, A., & Lauvås, P. (2011). Lavt sensorsamsvar – kan det bedres? [Low correspondence between examiners – can it be improved?]. *Uniped, 34*(4), 4–16.

Bonsaksen, T., Kvarsnes, H., & Dahl, M. (2016). Who wants to go to occupational therapy school? Characteristics of Norwegian occupational therapy students. *Scandinavian Journal of Occupational Therapy, 23*(4), 297–303. DOI: http://dx.doi.org/10.3109/11038128.2015.1105293.

Boud, D., & Falchikov, N. (Red.) (2007). *Rethinking assessment in higher education. Learning for the longer term*. Oxon, UK: Routledge.

Calvert, B., & Casey, B. (2004). Supporting and assessing dissertations and practical projects in media studies degrees: Towards collaborative learning. *Art Design & Communication in Higher Education, 3*(1), 47–60. DOI: http://dx.doi.org/10.1386/adch.3.1.47/0.

Grainger, P., Purnell, K., & Zipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment & Evaluation in Higher Education, 33*(2), 133–142. DOI: http://dx.doi.org/10.1080/02602930601125681.

Hand, L., & Clewes, D. (2000). Marking the difference: An investigation of criteria used for assessing undergraduate dissertations in a business school. *Assessment & Evaluation in Higher Education, 25*(1), 5–21. DOI: http://dx.doi.org/10.1080/713611416.

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395–412. DOI: http://dx.doi.org/10.1080/03075070410001682646.

Larsen, S., Johnsen, B. H., & Pallesen, S. (2006). Er opptaket til profesjonsstudiet i psykologi reliabelt? *Tidsskrift for Norsk Psykologforening, 43,* 221–225.

Lauvås, P., & Jakobsen, A. (2002). Hvilke krav skal eksamen tilfredsstille? In P. Lauvås & A. Jakobsen (Eds.), *Exit eksamen – eller?* Oslo: Cappelen Akademisk.

Lomas, L., & Nicholls, G. (2005). Enhancing teaching quality through peer review of teaching. *Quality in Higher Education, 11*(2), 137–149. DOI: http://dx.doi.org/10.1080/13538320500175118.

Newstead, S. (2002). Examining the examiners: Why are we so bad at assessing students? *Psychology Learning and Teaching, 2*(2), 70–75. DOI: http://dx.doi.org/10.2304/plat.2002.2.2.70.

Oslo and Akershus University College of Applied Sciences. (2011). *Bachelor programme in occupational therapy.* Oslo: Oslo and Akershus University College of Applied Sciences.

Rasch, B. E., & Eriksen, S. K. (2008). *En eller to sensorer? Et eksperiment.* Institutt for statsvitenskap, Universitetet i Oslo. Oslo.

Sadler, D. R. (1989). Formative assessment in the design of instructional systems. *Instructional Science, 18*(2), 119–144. DOI: http://dx.doi.org/10.1007/BF00117714.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education, 30*(2), 175–194. DOI: http://dx.doi.org/10.1080/0260293042000264262.

Sadler, D. R. (2013). Assuring academic achievement standards: From moderation to calibration. *Assessment in education: Principles, policy & practice, 20*(1), 5–19. DOI: http://dx.doi.org/10.1080/0969594X.2012.714742.

Saunders, M., & Davies, S. (1998). The use of assessment criteria to ensure consistency of marking. *Quality Assurance in Education, 6*(3), 162–171. DOI: http://dx.doi.org/10.1108/09684889810220465.

Scriven, M. (1991). Beyond formative and summative evaluation. In M. W. Maclaughlin & D. C. Phillips (Eds.), *Evaluation and education: At a quarter century* (pp. 19–64). Chicago, Il.: University of Chicago Press.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 85*(2), 420–428.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales – a practical guide to their development and use* (4 ed.). Oxford: Oxford University Press.

The Norwegian Association of Higher Education Institutions. (2011). *The grading system – general, qualitative descriptions.* Retrieved from http://www.uhr.no/documents/Karaktersystemet_generelle_kvalitative_beskrivelser.pdf

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*(1), 3–14. DOI: http://dx.doi.org/10.1016/j.stueduc.2011.03.001.

Wolf, H. (2004). Assessment criteria: Reflections on current practices. *Assessment & Evaluation in Higher Education, 29*(4), 480–493. DOI: http://dx.doi.org/10.1080/02602930310001689046.