Higher Education Studies; Vol. 6, No. 4; 2016
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Students' and Teacher's Experiences of the Validity and Reliability of Assessment in a Bioscience Course

Milla Räisänen[1], Tarja Tuononen[1], Liisa Postareff[1], Telle Hailikari[1] & Viivi Virtanen[2]

[1] Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland

[2] Faculty of Biological and Environmental Sciences, University of Helsinki, Helsinki, Finland

Correspondence: Milla Räisänen, Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland. Tel: 358-503-17-5465. E-mail: milla.raisanen@helsinki.fi

Received: September 1, 2016 Accepted: September 5, 2016 Online Published: November 24, 2016

doi:10.5539/hes.v6n4p181 URL: http://dx.doi.org/10.5539/hes.v6n4p181

Abstract

This case study explores the assessment of students' learning outcomes in a second-year lecture course in biosciences. The aim is to deeply explore the teacher's and the students' experiences of the validity and reliability of assessment and to compare those perspectives. The data were collected through stimulated recall interviews. The results showed that grades did not always reflect the learning outcomes and that the intended level of understanding was not always measured. In addition, the teacher and the students thought that the assessment criteria were unclear, which in turn led to the unreliability of the assessment. These problems with the validity and reliability of assessment led to perceptions that the assessment was unfair. The results imply that grades should be critically evaluated as indicators of the quality of learning outcomes. In addition, practical implications are discussed.

Keywords: assessment, grade, higher education, reliability, validity

1. Introduction

Assessment has an important role in higher education because it affects students' studying and the quality of learning outcomes. Course grades are used as objective measurements of student achievement. The grades are trusted and relied on for important decisions and they play a major role in students' lives. In addition, assessment might have a profound impact on students' sense of their own capacities and achievements (Sadler, 2009). Therefore, assessment should be valid and reliable. The validity of assessment refers to "grade integrity" which is about the extent to which grades correspond with the quality, breadth and depth of students' academic achievement (Sadler, 2009). For example, if the aim is to assess students' understanding, the exam should be designed to measure understanding instead of repetition of knowledge. Furthermore, in order to be valid assessment should be focused on measuring understanding of the core contents and the goals of the course (Birenbaum, 1996; Sadler, 2009). In addition, assessment should not be based on coincidences in order to be reliable (Sadler, 2009). The result should be the same regardless of who the assessor is or under which conditions the assessment is done. However, perfect accuracy in assessment is usually impossible because the judgements are based on subjective observations (Sadler, 2009).

Research shows that teachers may often experience that assessment is a separate part from teaching (Parpala & Lindblom-Ylänne, 2007). There is evidence that teachers do not necessarily have the competencies to assess students' learning in a valid and reliable way (Postareff, Virtanen, Katajavuori, & Lindblom-Ylänne, 2012). Research also shows that teaching and assessment practices change slowly (Deneen & Boud, 2014; Postareff, Lindblom-Ylänne, & Nevgi, 2007). Problems in assessment practices may be reflected in the reliability and validity of assessment. In the case of validity, grades do not always mirror the quality of learning. Other researchers have argued that students may successfully complete their courses without gaining a firm understanding of fundamental ideas (Ramsden, 2003; Struyven, Dochy, & Janssens, 2005). In addition, the teacher and students may have incompatible views about the level of knowledge being assessed. When Lingard, Minasian-Batmanian, Vella, Cathers, and Gonzalez (2009) studied the perceptions of staff and students studying biochemistry and physics regarding the level of difficulty of multiple-choice questions, they found that almost half of the students did notrecognise the level of assessed knowledge which was required from them. Therefore, they may have focused on lower-level knowledge in studying than what was required in the exam.There is also

evidence that teachers may compare students' answers to each other and change the grades afterwards in order to differentiate students' performanceor they may "grade on the curve" so that different grades are evenly distributed (Bloxham, Boyd, & Orr, 2011). In bioscience, the structure and nature of knowledge is cumulative and hierarchical, which is reflected in learning, teaching and assessment. It is essential for students to construct a knowledge base right from the start of their studies (Neumann, Parry, & Becher, 2002). Hard sciences contain a lot of factual and precise knowledge which may easily lead to the testing of such facts in exams (Lindblom-Ylänne, Trigwell, Nevgi, & Ashwin, 2006; Postareff et al., 2012). However, even in hard sciences the focus should be on assessing procedural knowledge instead of declarative knowledge (see, e.g., Odom & Bell, 2011).

In order to improve the reliability and validity of assessment, it should be based on clear and established assessment criteria. This is important because previous research clearly indicates that students adjust their learning according to what they think will be assessed (Biggs, 2003; Brown, Bull, & Pendlebury, 1997; Handley & Williams, 2011). Thus, to enhance student learning the assessment should be aligned with intended learning outcomes (Biggs, 2003). It is also important that the assessment is transparent (Segers, Dochy, & Gijbels, 2010) and that students are aware of the assessment criteria for different grades (Prosser, 2014; Sadler, 2005; Handley & Williams, 2011). Clearly set out criteria help also teachers to assess students' answers in a valid and reliable way (Yorke, Bridges, & Woolf, 2000). The assessment should also be authentic in such a way that knowledge and skills which are needed in realistic contexts are assessed (Segers et al., 2010).

Research combining a teacher's and students' experiences of the assessment, especially concerning its reliability and validity is rare. In the present case study, we focus on analysing assessment in one bioscience course. The aim is to deeply explore the teacher's and the students' experiences of the validity and reliability of assessment and to compare those perspectives. A recent study (Halinen, Ruohoniemi, Katajavuori, & Virtanen, 2014) indicated that there are needs for the development of assessment practicesin the study context, in general. Anotherstudy (Asikainen, Parpala, Virtanen, & Lindblom-Ylänne, 2013) showed that a year before, on this specific course, students who repeated knowledge, succeeded better in the exam than students who tried to understand. This indicates that there are some problems with the assessment in the course. This led us to explore the assessment in depth. A case study offers an opportunity to examine assessment from several perspectives in this specific course (e.g., Yin, 1994).

2. Method

2.1 Study Context

The data of this case study were collected from one Bachelor-level course in biosciences in collaboration with the teacher and the faculty. The course included lectures and a final exam. No other assessment than the final examination was included in the course. Each year approximately 100 students participate in the course. Most of the students were second-year students who had biosciences as their major. Based on the previous findings (Asikainen et al., 2013) we hypothesized that there are problems with the validity and reliability in assessment of the specific course.

The assessment method of the course was a "paper and pencil" exam at the end of the course. The exam was assessed on a scale from 1 to 5 (passed-excellent) which comprised the whole course grade. To pass the exam students must receive half of the points available. For the exam, students had to read lecture materials and they could also read a book which the teacher had recommended. In the exam, there were four questions and students had to answer three of them. All questions were essay tasks. In the first and second question, students were asked to describe the specific concept. In the third question, students were asked to compare different concepts. In the fourth question, students were given some facts about a specific research result and, based on this knowledge they had to explain how the research had been conducted. The aim behind this question was to test the application of knowledge.

2.2 Participants

The participants of the study were the teacher and four students of the course. The teacher of the course had a lot of teaching experience and had taught the course for many years. However, he had participated in a few short pedagogical training sessions. Formal pedagogical training is not a requirement in the university, although it is more or less appreciated nowadays. Among the students who volunteered for interviews, we selected four students with different grades. Another criterion for selecting the four students was that the teacher of the course was asked to explain the grading of these students' exam papers in the interview. All the students were second-year, bioscience students. Three of the students were female and one was a male student. Ages varied between 20 and 32 years.

2.3 Instrument

The students were informed about the study in the lectures. They were given interview invitations in a paper form and asked to give their contact information if they wished to participate in the study. Volunteers were then contacted by e-mail and interviewed after the course. The procedures followed the guidelines of the national research ethics committee (Academy of Finland, 2003).

In the interviews, the Stimulated Recall (SR) method was applied (see Lyle, 2003). The teacher and the students were interviewed after the exam papers had been graded by the teacher. The students' exam papers were at hand in the teacher and student interviews and the intention was to get them to recall the exam and assessment as well as possible. The student read his/her exam paper in the stimulated recall interview and the teacher went through those students' exam papers who participated in the interviews.

The teacher's interview focused on the assessment in the course but questions concerning teaching were also asked. The teacher was asked to think about the whole assessment process by several questions concerning the assessment seeking for to recall teacher's thinking while working on the assessment. The teacher was asked why he had asked the specific exam questions as well as what kind of knowledge and understanding were required in the exam. He was also asked to recall how he had assessed students' learning outcomesin each exam answer. In addition, the teacher was asked what kind of assessment criteria he had for different grades and why he had given specific points for specific answers.

In the interview, students were asked to read their exam answers. They were also able to see the comments made by the teacher. The interviews focused on students' experiences of assessment but questions related to students' learning and studying during the course were also asked. Concerning the assessment students were asked how they had prepared for the exam as well as what the teacher had told the students about the exam beforehand and how it affected their studying. In addition, the students were asked to recall and explain how they had written their exam answers in each question, i.e., if they had drafted the answer before writing and what they had been thinking while writing the answers. They were also asked to analyse what kind of knowledge and understanding were required in each question. In addition, they were asked what grade they had expected and what they thought about the fairness of the assessment. Moreover, clarifying questions were asked so that the students' answers could be correctly interpreted.

2.4 Analyses

The interviews were taped and transcribed. The interview data were analysed using inductive content analysis by the first two authors (Flick, 2002). Both authors analysed all the data but the first author analysed the students' interviews and the second author the teacher's interview in more detail. All descriptions which reflected aspects of assessment were analysed and grouped under different categories representing qualitatively similar descriptions. These categories represented aspects of the validity and reliability of the assessment. To ascertain the reliability of the analysis, the categorisation of the first author was compared to the categorisation of the second author. After that, the categorisation was negotiated together with all the authors. Agreement on this categorisation between the authors was high; all the authors identified similar main categories. Furthermore, all unclear cases were analysed together. In order to achieve a broader picture of assessment, descriptions in which students expressed how they prepared for the exam were analysed.

3. Results

The aim of this study was to construct a comprehensive picture of the assessment in that one specific bioscience course. In addition, the aim was to explore the teacher's and the students' experiences of the validity and reliability of assessment and to compare them. The main results were that both the teacher and the students experienced that assessment was not always valid and reliable. The grades did not always reflect the students' learning outcomes. In addition, there were problems in alignment of the assessment and clearness of the assessment criteria. For these reasons, assessment was not always perceived to have been fair. The main results are shown in Table 1.

Table 1. The students' and teacher's experiences of the assessment

Student (grade)	Experiences of the validity of assessment: Correspondence between grade and learning outcomes		Experiences of the reliability of assessment: Clearness of assessment criteria		Experiences of the fairness of assessment	
	Student	Teacher	Student	Teacher	Student	Teacher
Sarah (5)	rather good	good	unclear	clear	unfair	fair
Anne (4)	rather good	poor, (teacher experienced that the grade was too good)	unclear	unclear	unfair	unfair
Matt (3)	poor, (student expected a better grade)	poor, (teacher experienced that the grade was too good)	unclear	unclear	unfair	unfair
Maria (2)	poor, (student expected a better grade)	poor, (teacher experienced that the grade should have been better)	unclear	unclear	unfair	unfair

3.1 Teacher's and Students' Experiences of the Validity of Assessment

The teacher's and the students' experiences of the validity were related to the correspondence between grades and learning outcomes and the differences between intended and assessed knowledge.

3.1.1 Poor Correspondence between Grades and Learning Outcomes

As Table 1 shows, there was variation in how the teacher and the students experienced the correspondence between the grade and learning outcomes. The teacher experienced that the correspondence between grades and learning outcomes was good in only one case and was poor in other cases. None of the students experienced that the grades reflected their experienced learning outcomes, completely. Two students perceived that the correspondence was rather good; whereas two students experienced that the correspondence was poor. Next, the results and the students will be presented in more detail.

Sarah received a grade 5 (excellent) from the exam. She felt that the grade reflected her learning outcomes ratherwell. She had expected a good grade because she had studied enthusiastically and prepared well for the exam. The teacher also perceived that the grade mirrored the student's learning outcomes quite well. However, he revealed that there were contradictions in the student's answers. On the one hand, the student showed that she knew more than was required and had written more than was expected. On the other hand, the teacher experienced that the answers were not written in her own words. In the following example the teacher describes Sarah's answer:

> This is a good answer and I think that she knows more than is shown here … But in a way it's almost straight from the text. (Teacher)

Anne achieved a grade 4 (very good). The student's and the teacher's experiences of the correspondence between the grade and learning outcomes were contradictory. The student perceived that the grade reflected her experienced learning outcomes rather well. She had expected the grade because she had studied regularly during the course and prepared well for the exam. However, the teacher perceived that the grade did not mirror the student's learning outcomes. The teacher noticed that the answers were not as good as he had thought at first. In the following example, the teacher describes the student's answer:

> In this answer [answer two] there are no details at all and there are actually mistakes. She has got too many points from this … She should have got a lower grade. (Teacher)

Matt received a grade 3 (good) in the exam. Both the teacher and the student experienced that the correspondence between the grade and experienced learning outcomes was poor. However, their experiences were different. Matt expected a better grade because he thought that he had understood the assessed knowledge. The teacher perceived that the grade was too good when he compared Matt's exam answer with Maria's exam answer. When the teacher read these students' exam answers, he noticed that Matt's answers were not as good as the grade suggested. The teacher was not sure, if Matt had understood the subject and that a few important

aspects were missing in the answers. In the interview, the teacher also described the correspondence between the grades and learning outcomes in more general terms:

> The grades should reflect it [learning outcomes] but unfortunately they aren't reflecting it in this exam because if you answer these tasks, you don't have to know anything else but what has been said in the lectures and you can achieve a good grade. (Teacher)

Maria received a grade 2 (satisfactory). Both the teacher and the student perceived that the correspondence between the grade and the experienced learning outcomes was poor. Maria was disappointed because she had expected a better grade and she thought that she had understood the main points of the course. She was also quite worried about her learning in general because she had studied the same way in the course as usual. In the following example, the student describes the correspondence between the grade and her learning outcomes:

> It's a little bit more [learning outcomes] than what the grade reflects. That course didn't go perfect but I didn't expect such a poor grade. (Maria)

The teacher also perceived that the grade did not reflect this student's learning outcomes when he read Maria's answers again. The teacher was quite upset when he noticed that.

> Oh my God (sigh). Yes, this is actually quite a good [answer]. As I said, this should be raised to a three [grade] ... Oh (sigh), I must have been in a bad mood when I read this. (Teacher)

The teacher noticed that students could achieve good grades just by memorising what was written in the course materials or what the teacher had said in the lectures. In addition, this course was especially problematic because the students knew that almost the same exam questions are asked in the exam every year and thus they could study them beforehand.

3.1.2 Differences between Intended and Assessed Level of Understanding

Problems with validity of the assessment were also revealed as the exam questions did not always measure what they intended to measure. Both the teacher and the students experienced that the exam did not always require the intended level of understanding. In addition, the level of understanding was not always taken into account in the assessment criteria and the aims of the course were not applied in the assessment.

The teacher noticed that comparing of different concepts was not required in the exam although he had intended to measure it. In the following example, the teacher describes the comparing task:

> In this comparing question, if you just remember the slides I had shown in the lectures, you can actually manage the task quite well. You don't actually have to understand these things. (Teacher)

The teacher noticed that the task in which students were asked to compare different concepts did not require real comparison because students could get good points just by reproducing how the teacher had compared the concepts in the lectures. Students also wondered what kind of knowledge was required in the task. In the following example, Matt describes the comparing task and wondered why he had got the highest mark for the task without really comparing the concepts:

> I actually didn't compare these concepts that much. It has obviously been enough that I've just written about these techniques one after another. (Matt)

In addition, the teacher had intended to measure the application of knowledge, which was also one of the aims of the course. However, the teacher noticed that the task, in which students were asked to apply knowledge, did not actually require the applying of any knowledge and he was confused when he realised that the exam mostly required just a reproduction of knowledge:

> You just have to remember things. For God's sake (surprised), you don't have to apply and integrate knowledge in any of these tasks. In the fourth task, the students should apply knowledge, but actually it's enough that you just remember and understand what has been talked about. (Teacher)

3.2 Teacher's and Students' Experiences of the Reliability of Assessment

The experiences of the reliability of the assessment were related to unclear assessment criteria. Both the teacher and the students experienced that the assessment criteria were unclear.

3.2.1 Unclear Assessment Criteria

The teacher did not have clear assessment criteria when assessing the students' learning, which caused some problems for his assessment. In the interview, the teacher mentioned some of the assessment criteria he used for

the assessment, but it was shown that the teacher used different assessment criteria when assessing students' answers.

The students were aware that there would be four questions in the exam and that they had to answer three of them. As Table 1 indicates, all the students experienced that the assessment criteria were unclear. Anne described the unclearness of the assessment criteria in the following way:

> I'd like to know why I got only 3.5 points for the first task and I don't understand why the teacher had underlined this sentence in particular. (Anne)

In the following example, the teacher describes Anne's answer:

> In this answer I think I've been too kind in my evaluation or maybe I just haven't noticed that in the beginning of the answer the student hasn't really understood what she was supposed to … Please, comfort me. This student has definitely got too many points. (Teacher)

This example shows that the assessment criteria were unclear for the teacher and he wondered how he had assessed the answer. The teacher noticed the same kind of problem in other students' answers. He had given either too many or too few points. In addition, the teacher noticed that he had given a similar number of points for answers that were different in quality. He was surprised and confused when he noticed this. In the following example, the teacher describes the assessment of Maria's and Matt's answers:

> These answers differ like day and night in every way although I've given the same points for them … Oh my God, this student [Matt] should have had a lower grade. (Teacher)

3.2.2 Experiences of Unfairness of the Assessment

The results of this study show that both the teacher and the students experienced that assessment was not always fair. Experiences of unfairness were related to problems with the reliability and validity of assessment.

In Maria's case, experiences of unfairness were shown clearly. Maria received a much worse grade from the exam than she had expected. In the following example, Maria describes assessment in one task:

> I got only one and a half points from the second task. There was one bigger mistake in the middle. I feel that I got points only from what I said before the mistake. Everything I have written after the mistake has been ignored somehow. (Maria)

This example shows that Maria did not know how the teacher had assessed her answer or why she had received so few points from the task. In addition, she felt that there was something unfair in the assessment. In the interview, Maria said that receiving a grade 2 was depressing and she had begun to doubt her abilities as a student.

The teacher also noticed that his assessment had not been fair when he compared Maria's answer to another student's answer. He emphasised that Maria should have received a better grade.

> One thing I can say for sure, these papers haven't been assessed one after another (sigh), she should have had a higher grade. (Teacher)

During the interview, the teacher recognised many problems in the assessment and he realised that he would have to change the assessment. The teacher also described different things which might affect the reliability of the assessment in more general terms:

> That is actually quite scary, the evaluation of the answers can vary two points depending on the time of the day, the mood or the order in which I've assessed the answers. (Teacher)

4. Discussion

Our results revealed severe and worrying challenges for the assessment of students' learning outcomes. Our main finding was that the assessment of students' learning was not always valid and reliable. Thus, our hypothesis based on the previous study (Asikainen et al., 2013) was confirmed. Asikainen et al. (2013) showed that students, who sought deep understanding, received a lower grade than students who repeated knowledge. Our previous research on assessment shows that the students and teachers did not always experience the assessment as reliable and valid (Hailikari, Postareff, Tuononen, Räisänen, & Lindblom-Ylänne, 2014).

A major concern in terms of validity was that the grades did not always mirror the learning outcomes. In other words, grades did not reflect the quality of students' achievement very well. This has also been shown in earlier research in this specific course (Asikainen et al., 2013). The results imply that grades do not necessarily reflect the quality of learning. This study confirms the role of assessment for students' studying and learning (Biggs,

2003; Brown et al., 1997). In this course, an excellent grade could be received by route memorisation without necessarily any real understanding because the exam mostly required the reproduction of knowledge. Vice versa, students could achieve a poor grade even though they understood the assessed subjects.

The other problem concerning validity was that the teacher assessed different level of understanding which was intended to. Thus, the teacher's intention in assessment was in higher level than what was implemented in practice. This problem is related to the lack of alignment in assessment. A previous study has proved that the validity of the assessment is reduced if the given tasks are irrelevant (Sadler, 2009).

Concerning the reliability of the assessment the results revealed that the assessment criteria were unclear both for the students and the teacher. Previous studies have indicated that students are not always aware of the assessment criteria and what is required from them (Rust, Price, & O'Donovan, 2003; Sadler, 2009). Our results underline the importance of clear assessment criteria. If the assessment criteria are not set out clearly beforehand, they can vary during the actual process of assessment and even from one student to another (Yorke et al., 2000), which was also revealed in this study. In our study the teacher compared the students' answers to each other, which also reduced the reliability of the assessment. In addition, this study confirms earlier findings that many internal and external factors affect assessment and that assessment is always subjective (Sadler, 2009). These problems with the validity and reliability in assessment caused students' experiences of unfairness. Students were not always aware of how they had got their points during the exam. However, they still considered the assessment to be reliable and fair. It became clear that the students trusted the assessment conducted by their teacher and did not question the grading.

Previous research has shown that a teacher and students may experience the assessment in a different way (Lingard et al., 2009). This was shown also in the present study. Most of the differences concerned the validity of assessment, in more detail, experiences of the relationship between grades and learning outcomes. Furthermore, there were also some differences in the teacher's and the students' experiences of the reliability of assessment and they focused on unclearness of assessment criteria. In addition, the teacher's and the students' experiences of the fairness of assessment differed.

The aim of the study was to construct a comprehensive picture of the assessment in one bioscience course. This study was a case-study with only a few participants: just four students and one teacher on one course were interviewed. For this reason, the results cannot be directly generalised to other courses and other disciplines. It would be even possible that the findings were the result of poor design of the assessment criteria in only this particular course reflecting the teacher's lack of pedagogical awareness or pedagogical training. Hence, generalizing the results to other situations should be done with caution. However, other results from the same context have argued that assessment practices at the course level should be considered, while a variety of pedagogical support is needed in the development of high quality assessment (Halinen et al., 2014). In addition, the cumulative and hierarchical nature of knowledge and a tendency to assess declarative knowledge instead of procedural knowledge might lead to similar problems in other biosciences courses as well.

The use of stimulated recall interviews was essential for obtaining knowledge that would not have been obtained otherwise. By using this method, contradictions between the teacher's intentions and practice were revealed. In addition, it was easier for students to assess the reliability and validity of the assessment when they had their own exam papers to hand. Furthermore, the students were able to recall emotions before and after assessment.

Previous research has shown that assessment strongly influences students' studying and learning (Biggs, 2003). Teachers should make students aware of course demands and assessment criteria in order to ensure that they share the same goals and understanding (Rust et al., 2003). Therefore, teachers' pedagogical awareness is essential in improving the reliability and validity of assessment. Assessment and teaching should help students to construct higher level of knowledge and thus assessment should focus on procedural knowledge instead of declarative knowledge. This may be done by using formative assessment and multiple assessment methods (e.g., Crisp, 2012; Mintzes, Wandersee, & Novak, 2001; Sneddon, Settle, & Triggs, 2001).

Our results indicated that grades do not necessarily reflect the quality of learning outcomes. We suggest that grades should be critically evaluated as indicators of learning outcomes, especially, when more than the mere repetition of knowledge is expected from students. It is striking that in the specific course both the teacher and the students found problems in reliability, validity, and fairness when these should be the basic components of any high quality assessment. Further research, with larger sample sizes and in variable courses and disciplines, is needed in order to explore if grades are reliable indicators of learning outcomes. Our findings suggest that it is important to enhance university teachers' pedagogical awareness regarding the reliability and validity of assessment and the role of asses learning in order to ensure high quality assessment practices. If teachers are not

aware of these aspects of assessment, it is difficult for them to change their assessment practices. The awareness of assessment can be enhanced through pedagogical training and through using more collaborative practices (Deneen & Boud, 2014). In addition, it is important to raise students' awareness of the assessment processes and provide them possibilities to engage in the assessment process (Boud & Falchikov, 2006). Thus, it could have a positive influence on students' experiences of the reliability and validity of assessment.

References

Academy of Finland, Guidelines on Research Ethics. (2003). *Helsinki: Academy of Finland*.

Asikainen, H., Parpala, A., Virtanen, V., & Lindblom-Ylänne, S. (2013). The relationship between student learning process, study success and the nature of assessment: A qualitative study. *Studies in Educational Evaluation*, *36*(4), 211-217. http://dx.doi.org/10.1016/j.stueduc.2013.10.008

Biggs, J. (2003). Teaching for quality learning at university. In *What the student does* (2nd ed.). Buckingham: The Society for Research into Higher Education & Open University Press.

Birenbaum, M. (1996). Assessment 2000: Towards a Pluralistic Approach to Assessment. In M. Birenbaum, & F. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-29). Boston: Kluwer Academic Publishers. http://dx.doi.org/10.1007/978-94-011-0657-3_1

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. http://dx.doi.org/10.1080/03075071003777716

Brown, G., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, *31*(4), 399-413. http://dx.doi.org/10.1080/02602930600679050

Crisp, G. T. (2012). Integrative assessment: Reframing assessment practice for current and future learning. *Assessment & Evaluation*, *37*(1), 33-43. http://dx.doi.org/10.1080/02602938.2010.494234

Deneen, C., & Boud, D. (2014). Patterns of resistance in managing assessment change. *Assessment & Evaluation in Higher Education*, *39*(5), 577-591. http://dx.doi.org/10.1080/02602938.2013.859654

Flick, U. (2002). *An introduction to qualitative research* (2nd ed.). London: Sage Publications.

Hailikari, T., Postareff, L., Tuononen, T., Räisänen, M., & Lindblom-Ylänne, S. (2014). Students' and teachers' perceptions of fairness in assessment. In C. Kreber, C. Anderson, N. Entwistle, & J. McArthur (Eds.), *Advances and innovations in university assessment and feedback* (pp. 99-113). Edinburgh: Edinburgh University Press. http://dx.doi.org/10.3366/edinburgh/9780748694549.003.0006

Halinen, K., Ruohoniemi, M., Katajavuori, N., & Virtanen, V. (2014). Life science teachers' discourse on assessment: A valuable insight into the variable conceptions of assessment in higher education. *Journal of Biological Education*, *48*, 16-22. http://dx.doi.org/10.1080/00219266.2013.799082

Handley, K., & Williams, L. (2011). From copying to learning: Using exemplars to engage students with assessment criteria and feedback. *Assessment & Evaluation in Higher Education*, *36*(1), 95-108. http://dx.doi.org/10.1080/02602930903201669

Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context? *Studies in Higher Education*, *31*, 285-298. http://dx.doi.org/10.1080/03075070600680539

Lingard, J., Minasian-Batmanian, L., Vella, G., Cathers, I., & Gonzalez, C. (2009). Do students with well-aligned perceptions of question difficulty perform better? *Assessment & Evaluation in Higher Education*, *34*, 603-619. http://dx.doi.org/10.1080/02602930802287249

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. *British Educational Research Journal*, *29*, 861-878. http://dx.doi.org/10.1080/0141192032000137349

Mintzes, J., Wandesee, J., & Novak, J. (2001). Assessing understanding in biology. *Journal of Biological Education*, *35*, 118-124. http://dx.doi.org/10.1080/00219266.2001.9655759

Neumann, R., Parry, S., & Becher, T. (2002). Teaching and learning in their disciplinary contexts: A conceptual analysis. *Studies in Higher Education*, *27*, 405-417. http://dx.doi.org/10.1080/0307507022000011525

Odom, A., & Bell, C. (2011). Distinguishing among declarative, descriptive and causal questions to guide field investigations and student assessment. *Journal of Biological Education*, *45*, 222-228. http://dx.doi.org/10.1080/00219266.2010.549495

Parpala, A., & Lindblom-Ylänne, S. (2007). University teachers' conceptions of good teaching in the units of high-quality education. *Studies in Educational Evaluation*, *33*(3), 355-370. http://dx.doi.org/10.1016/j.stueduc.2007.07.009

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2007). The effect of pedagogical training on teaching in higher education. *Teaching and Teacher Education*, *23*, 557-571. http://dx.doi.org/10.1016/j.tate.2006.11.013

Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' conceptions of assessment and their assessment practices. *Studies in Educational Evaluation*, *38*(3-4), 84-92. http://dx.doi.org/10.1016/j.stueduc.2012.06.003

Prosser, M. (2014). Perceptions of assessment standards and student learning. In C. Kreber, C. Anderson, N. Entwistle, & J. McArthur (Eds.), *Advances and innovations in university assessment and feedback* (pp. 114-128). Edinburgh: Edinburgh University Press. http://dx.doi.org/10.3366/edinburgh/9780748694549.003.0007

Ramsden, P. (2003). *Learning to teach in higher education* (2nd ed.). London: Routledge.

Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education*, *28*, 147-164. http://dx.doi.org/10.1080/02602930301671

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

Sadler, R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*, 807-826. http://dx.doi.org/10.1080/03075070802706553

Segers, M., Dochy, F., & Gijbels, D. (2010). Impact of assessment on students' learning strategies and implications for judging assessment quality. In P. Peterson, E. Baker, & B. McGraw (Eds.), *International Encyclopedia of Education* (3rd ed.). Oxford: Elsevier. http://dx.doi.org/10.1016/B978-0-08-044894-7.01625-0

Sneddon, J., Settle, C., & Triggs, G. (2001). The effects of multimedia delivery and continual asssement on student academic performance on a level 1 undergraduate plant science module. *Journal of Biological Education*, *36*, 6-10. http://dx.doi.org/10.1080/00219266.2001.9655788

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, *30*, 325-341. http://dx.doi.org/10.1080/02602930500099102

Yin, R. K. (1994). Case study research. In *Design and methods*. California: Sage Publications.

Yorke, M., Bridges, P., & Woolf, H. (2000). Mark distributions and marking practices in UK higher education. Some challenging issues. *Active learning in higher education*, *1*, 7-27. http://dx.doi.org/10.1177/1469787400001001002