Self-evaluation

I am writing this self-evaluation after I have been reading the two other reflection assignments from my peer students. And if I had done it the other way round, my self-evaluation would be probably different (this could be perhaps classified under the halo effect). I realized that one other text mentioned the same difficulty as I in distinguishing between teaching and assessment. While the other text interpreted this difficulty positively, I took the other turn and focused on the burden for teachers and students by this "fusion".

Again: it would help my self-assessment if I would have an exemplar. Now I compare to two texts of unknown quality. I have a feeling, but as a scientist, I do not trust my feelings: it is all about measuring). I have already once asked for an exemplar, but these are difficult to find (to help others in the same situation, I probably will simply put up all my written musings for this course to my web pages for future reference by others, disregarding all embarrassment).

There are also technical issues: The text is too long. Max. was four pages and I only managed to stay within the limit by extensively minimizing the page borders and the font size for the bibliography. Ok, I managed to gather many references (but I have read barely half of them, even though I managed to find the full text for all of them and they are patiently waiting in the Pedagogy subsection of my Zotero bibliography management software).

Now a bit more about the content: I fail to mention several buzz words and the birds' view of the topic. The reason being that this text is cobbled together by pasting my lecture notes for each contact session, then adding some glue text and references and shortening by about 50%, running Grammarly over the text and voilà.
I don't even manage to mention the phrase "constructive alignment" once! Whether this is negative depends of course much on what this reflection is all about. I discuss at least some of the concepts (alignment, formative versus summative assessment and the fact that there are no clear demarcation lines). I mention what I might (or might not) implement in my future teaching, including some ideas about why or why not. I also definitely have changed my view, but I am not sure whether in a good direction. I, e.g. still have no clue about what is the consensus agreement about many topics (the "real" effects of assessment as learning and continuous assessment for students), and that shows in the text. Did I not pay attention during the course? For sure it was challenging: The last contact session (#3) was so stressful, that it launched a full-blown migraine attack, which nowadays I have extremely seldom (maybe twice a year), but pedagogy got it done.

Back to the content: I also touch topics that were not part of the course (e.g. AI in assessment). This is bad perhaps only because it takes space away from the already limited maximum page count.

Even though this does not belong to this text: The most important voice in this matter (i.e. the students' voice) was never heard during the course. Even though we were the students in this course, I think that this misses the point and is only relevant to the degree to which we can objectively remember our studies and we are certainly not a representative sample (I am not sure whether all of the participants can be classified as "Susans", but maybe that is not far from the reality).

When I review a manuscript in my field, it takes usually about one day of work to do it thoroughly. And often I ask a colleague for help or input. How on earth can I do a good job in a field where I am a novice and I have perhaps one hour's time to do it? Impossible.

UP 2.2 Assessment of Learning and Giving Feedback - Reflection Assignment

Introduction & caveats

The usefulness of reflective writing for assessment (like this one) is discussed in Biggs and Tang (2013), but they only use self-citations to support their claims, which might be an indication of less than average evidence for their usefulness. Even if we accept the usefulness of reflection as a given, it is arguable whether the adequate explicit formulation of such reflection in a written essay correlates well with the underlying cognitive process, since the quality of the essay is also determined by linguistic ability. In one study, it was shown that in evaluating reflective writing, much emphasis is given on the linguistic richness of the text, which biases the grading in favor of students that study language and also tends to favor native speakers: "A strong correlation was found between the overall reflective score and the total number of linguistic resources used, showing that reflective writing tends to be linguistically richer." (Birney, 2012). The other aspect of this assignment, i.e. to monitor my own development (new insights, planned changes to my courses, etc.) is also a bit hampered since - after contact session #2 - I am barely halfway through the course and (hopefully) the second half will provide as much material and inspiration as the first half. If not, shame on the teacher. If so, this assignment can only address half of the course's content. With this caveats in mind, here we go:

Alignment of the assessment with the intended learning outcomes (ILOs)

"Alignment" of the assessment with the ILOs seems to be to major take home message of the whole course. If the assessment does not measure the ILOs, students are encouraged to learn the wrong things (e.g. memorizing answers to multiple choice questions instead of understanding and critical appreciation). This makes intuitive sense and is not really new to me. Perhaps due to the close relationship that biomedical researchers have with the Impact Factor and its shortcomings, they know the concept of bad alignment (Goodhart's Law: "when a measure becomes a target, it ceases to be a good measure").

Assessment task (AT) versus Intended Learning Outcome (ILO) assessment

What was less clear to me was the difference between assessing the assessment task or the intended learning outcome. However, this seems to me a very important difference. If the students fails completely to address the task but shows that the intended learning outcome has been perfectly reached, then a maximum grade could be given. However, if the task is assessed, a very low grade might be given. If the teacher has not decided whether to assess the task or the ILOs (and if this is not transparent for the student), this could lead to very big surprises.

Explicitly, I became aware of this difference when looking at case 1 from task 10.1 of the pre-course reading - Chapter 10 of (Biggs and Tang, 2011). It is about this difference (and I modify my answer to this question): If the assessment was directed at the task, several of the answer options are possible, but if it was directed at the learning outcome, the actual question of the task is not the central issue. Instead, if the student's answer clearly shows that the intended learning outcomes have been reached, the grading can be done without the need to repeat the assignment, deduct points, etc.

Chaos everywhere (at least for me): Teaching, Learning or Assessment?

My (biased) opinion that the research field of pedagogy is very confusing became reinforced during the interactive sessions and when reading some of the course literature, There is hardly a research paper for which one cannot find another, that argues in the opposite direction. Thus, I very much support the request of one course participant for reviews that present the scientific consensus opinion. It is difficult to understand why there are no such systematic reviews comparable e.g. to the Cochrane Reviews or the British Medical Journal's Best Practice in medicine, which are cornerstones of evidence-based medicine for health care professionals. There is certainly an unmet need and an opportunity for a comparable resource for pedagogy professionals! There are "reviews" (e.g. (Black and Wiliam, 1998; Hattie and Timperley, 2007) , but these also contain own analysis and do not systematically compare and classify the strength of the evidence (i.e. they do not digest it well for practical applications).

Mixing, separating or eliminating summative assessment?

The borders between learning and assessment are blurred. In my opinion, meaningful summative assessment is only possible after learning has happened. However, formative assessment can happen nearly all along the learning process. In my opinion, this creates a tension that is difficult to resolve. Nevertheless, some authors even argue for the superiority of combined formative-summative assessment (Buchholtz et al., 2018). However, I do not follow their logic. E.g. in Buchholz et al (2018), there is a logical paradox: "In this paper, we argue that corresponding to the different aspects of teaching competence, different forms of assessment are beneficial when it comes to assessing the **outcomes** of pre-service teachers' learning processes." They aim to assess the outcomes, but formative assessment is about supporting the learning **process**. The outcome is the final result, whereas formative assessment happens during the learning process. The mixing of formative and summative assessment carries further burdens for the students: Constant

surveillance modifies behaviour (Penney, 2016). There are whole schools (of psychology) that have abolished summative assessment within the legally possible boundaries based on its presumed negative effects (e.g. Rudolf Steiner schools).

Assessment as Learning (AaL)

The notion of assessment as learning (AaL) is one intentional blurring between learning and assessment; another one is the relabelling of whole learning entities as assessment, e.g. when declaring a final-year project or a case-study an assessment format, as id done in chapter 12 of Biggs and Tang (Biggs and Tang, 2011). However, this blurring is sold by some as a positive thing (Dann, 2014). But again, there is considerable disagreement between different scholars on the current views on AaL (Brown, 2019). In my opinion, specifically the higher level functioning knowledge might not thrive under the pressure of constant assessment (even if the assessment is supportive) and there is evidence to support this (Amabile et al., 2002; Harb-Wu and Krumer, 2017).

Peer-assessment

Peer- and self assessment is what's happening in real life. Assessment in professional life happens mostly as peer- and self-assessment (at least in academic circles). Very pointedly, most of the groups in this UP2.2 course did forgo the peer assessment and decided to weight the group assignment equally, which supports the reality of the caveats that have been described in the research literature: refusal (Gibbs, 1999), stress (Pope, 2001) and the problematic of the Dunning-Kruger effect (Kruger and Dunning, 1999), which apparently works also in this context into both directions (Lejk and Wyvill, 2001)[1]. Whether the decisions in favour of equal grading for everybody were made despite better knowledge, or just reflect the fact that most participants need to get this course done on top of a 50-hour work week, is up for discussion. In any case, it is difficult to argue with Biggs and Tang (2011): "The common practice of simply awarding an overall grade for the outcome, which each student receives, fails on all counts." Whether group assignments make sense especially for such a time-constraint student clientele is questionable (who planned this course???). This typically leads to the minimization of workload by distributing tasks according to the individual strengths of the group members, which can produce a good result despite of minimized learning for each individual[2]. As possible grading scheme, Biggs and Tang mention the allocation of a fixed amount of points between group members, which introduces again the questionable aspect of competition. If Leonardo da Vinci, Einstein, Gauss and Riemann happen to end up in the same group, each of them would receive a totally inadequate amount of points.

Use of rubrics

Assessment criteria have to be clear for the students. It seems generally accepted that the use of descriptive rubrics to guide the grading process can be helpful both for students and (perhaps even more so) for teachers (Biggs and Tang, 2011), original publication by (Norton, 2004). However, when looking at some of the rubrics we received during the contact teaching, it seems to be a challenge to come up with distinct and clear description for each grade. "in between" was the most common description for some of these rubrics; this is not very helpful for students or teachers. Just a few months back, we updated our study program's rubrics for the grading of a Master's thesis, and we also left out the descriptions for grades 0, 2 and 4. Even if the rubrics are filled out nicely, they still use language which is subject to interpretation and ambiguity. And I still have to meet the student who has read them (they are virtually impossible to find; try and tell me the URL!). We actually has several debates about these and we removed several criteria which were from a legal point of view very difficult to maintain (e.g. the requirement of originality for the best grade; the consensus opinion was that originality can easily be introduced into a natural science MSc thesis, but it is easy to make up an original, but worthless interpretation for some data ("aliens visited and exchanged the test subjects during the night").

Rubrics alone do not help if they are not equally interpreted and applied

Biggs and Tang are quite outspoken about the need to validate the teacher's ability to apply the criteria correctly: "The reliability of their interpretations of the criteria by each may be tested by assessing a sample of the same scripts and repeating this procedure until they reach a high degree of consensus, say of the order of 90% within a range, say, of ±1 grade." Such validation I have never seen happen in reality. I have some experience from grading the applications for study places in our MSc programme (TRANSMED), which is done in our programme independently by three evaluators. And it is surprising how different the free-form answers to the questions are graded by different evaluators.

Is this a bad or a good thing? Internally we have agreed that it is a good thing and haven't even tried to address these differences. Different evaluators have different viewpoints and put emphasis on different aspects and thus we ensure diversity among the students. Only when the differences are drastic, we discuss individual cases. However, we are

[1] The fact that Johnston & Miles did not see the Dunning-Kruger effect was the original reason why I started to look at their calculation, which appears to be flawed.

[2] Just as a side note: In most fields, minimization of work load is a goal (e.g. in mathematics the shorter way of calculation or the shorter proof is better, in informatics the shorter code is better). However, when reading some of this course literature, I conclude that pedagogy is not among those fields that reward brevity.

dealing here with a special case of diagnostic assessment and for summative assessment, some kind of validation would be nice, but I doubt this idea would fly in our faculty…

AI versus humans in assessment

Artificial intelligence has been already used to replace assessment, using the same logic: "Inconsistency occurs when raters are either judging erratically, or along different dimensions, because of their different understandings and interpretations of the rubric" (Zhang, 2013). Also Biggs and Tang report that "Sophisticated computer assessment" can support or replace human grading, but the term "sophisticated computer assessment" is not very informative (they likely mean AI, and more specifically neural networks/deep learning).

Issues with human raters are the halo effect (when a single attribute of a performance is generalized to the whole performance), stereotyping (especially when the rater knows the student), perception difference (undervalueing a performance after just assessing an exceptionally good performance), rater drift (when the rating shifts over time).

Issues with current AI ("deep learning") is that they are trained with human samples (e.g. students' assays and the corresponding teachers' gradings) and therefore they reflect the sum of the human biases. Such biases in AI systems have surfaced in many other areas where AI has been used to evaluate individuals like job applications, face recognition, medical systems (Gomez and Rosenberg, 2019; Harwell, 2019; Obermeyer et al., 2019). Even though I am very confident about AI's capability to perform the task, I am as skeptical about the correct training of such systems and about the limited understanding of the true underlying principles that these neural networks establish (Turunen, 2019). AI (and even more so proprietary AI) is a black box that often uses sets of characteristics completely different from humans to arrive at their conclusions, which can lead to catastrophic errors (e.g. recognizing a flying cow as an airplane as opposed to a cow, because the AI was trained on images of cows (which are of course all earth-bound). Similar errors could probably happen in AI assessment.

Actionable learning goals

So far, I only assess pass/fail in my courses. I proclaimed in the first pre-course assignment that I have not much need for assessment skills, but that is not entirely correct. I happened to get two assessment assignments during the last weeks. These assessments are, however, only diagnostic:

- I need to assess the applications to our study program (TRANSMED).

- My yearly course (https://courses.helsinki.fi/en/dpbm-135/131042336) happened to be for the first time massively oversubscribed. Hence, I needed to assess students, deciding who would be accepted and who not. These tasks appear to me still totally different from everything we have discussed during the course so far. Even diagnostic assessment at the beginning of a course (as done for this course) is different as the students that are not accepted into the study programme are simply removed from the pool and their assessment results do not influence the teaching and the diagnostic results are never used to monitor progress. It seems to me that there needs to be another type of assessment (or a subtype to the diagnostic assessment), which could be called "eliminatory" assessment. These types of assessment are actually very common (e.g. job interviews), but e.g. there is no possibility for alignment as the learning/teaching and assessment are executed mostly by completely different entities.

I am not going to change grading from pass/fail to a scale. Mostly, I want to implement a better formative assessment, and I am looking at an online implementation of the ordered-outcome format to be done by the students over the course period to make them realize when they have not grasped core concepts. However, my yearly major course starts 2.12. (full-day lab course until 17.12., colliding with one UP2.2 contact session) and there is no time (especially since my lab runs on 50% of its usual strength due all three female lab members being on maternity leave).

Practical improvements need to be supported by the system

Some very practical matters are useful for me, e.g. that instead of assessing student by student, I should assess question by question in order to avoid the halo effect[3]. However, such a strategy is made difficult by the way how the evaluation procedure is implemented at our faculty. I get every student's answers as a single PDF file. Evaluating question by question would mean to cycle between 140 open PDF files. There is clearly room for the improvement of the system, which is certainly going to happen very soon (just kidding). Also other very useful hints are almost impossible to implement in our study programmes current system: We do not use blinding (I know the students' names and their previous performances), and we also do not randomize the documents like Bing and Tang suggest (Chapter 11, page 233), which makes only sense when evaluating question-by-question as opposed to student-by-student.

Some other ideas (i.e. the ordered-outcome format, Biggs and Tang, Chapter 11 page 235) are nice, but need effort to implement. I am just trying to come up with some ordered-outcome quiz for my present course, but the difficulty with

[3] Strangely, when discussing the assessment of portfolios, Biggs and Tang argue for holistic assessment, which would be subject to a stronger halo effect (one very good item influencing the portfolio's overall assessment).

coming up with appropriate quizzes seems to increase with the level of teaching (easy for fresh undergraduates, but very difficult for advanced PhD students), and it is obviously lots of more work compared to the typical MCQ.

The *Extreme Apprenticeship* concept sounds intriguing and worth trying, and I will try this fully with next year's course (with this years course, I might make some small steps into this direction). The inversion with students doing the tasks first and then having the lecture sounds very appealing to me. It's like learning by trying things out, which we call *research* or *science*, because we do not know the answer (Rämö et al., 2019; Vihavainen et al., 2011). Often we do not even know how to get to the answer. We try this and that, and slowly approximate our way towards an idea or concept. Only afterwards we streamline our actions when we write the paper (which is of course entirely fictional as it rarely describes all the wrong turns we took before we figured out what to do in order for things to work out).

Miscellaneous ramblings

Pedagogy feels to me like medicine practiced 100 years ago. Many different ideas were tested (most of which were abandoned) and there was no clear framework as today, where evidence-based medicine is the accepted framework and science-based medicine perhaps the next step ahead (adding the mechanistic explanation as a prerequisite for improvement). In medicine, expert opinion (consensus best practice) is gathered in a systematic way and offered to the practitioners. Citing (quite freely according to my memory) our course teacher: "One of the goals of this course is that you become aware of different approaches and that you can try out different things, being aware of all the differences between your situation and the situation that has been described in a particular publication." Being cynical, one could add: "Let's try electroshock, lobotomy and blood-letting. It might or might not work."

A question during the contact session on Nov. 12th was about such consensus opinions (review articles), but it seems that there is not much out there. Formulations like "I am not going to tell you", "it can work", "you can try this and that, being aware of all the differences" etc. are indicative that there is a lack of systematic analysis. Is this lack is due to the fact that less research is done (as opposed to medicine), is there less funding?

Despite the attempts to improve assessment, the literature acknowledges underlying problems (e.g. Biggs and Tang on page 239: "All this is fairly arbitrary, but then using numbers to quantify qualitative data always is.") If it was arbitrary, why not throw dice? Some research funding agencies have started to distribute a certain amount of their money based on a lottery system instead of having the applications assessed by "experts" (Adam, 2019; Gross and Bergstrom, 2019). Whether this is merely a reaction to the inefficiency of the assessment system or an acknowledgement of the impossibility of meaningful assessment is up for debate; likely it's a combination. Maybe universities should learn from these examples (at least at the admission stage)! Another idea to integrate randomness into assessment was to have students submit many reports, but randomly assess only a subset. This reduces the teacher's work, and apparently improved as well report quality (Biggs and Tang, 2011), original research by (Gibbs, 1999).

Literature

Adam, D. (2019). Science funders gamble on grant lotteries. Nature 575, 574–575.

Amabile, T., Hadley, C.N., and Kramer, S.J. (2002). Creativity Under the Gun. Harv. Bus. Rev.

Biggs, J., and Tang, C. (2011). Teaching for Quality Learning at University (Open University Press).

Birney, R. (2012). Reflective Writing: Quantitative Assessment and Identification of Linguistic Features. Waterford Institute of Technology.

Black, P., and Wiliam, D. (1998). Assessment and Classroom Learning. Assess. Educ. Princ. Policy Pract. *5*, 7–74.

Brown, G.T.L. (2019). Is Assessment for Learning Really Assessment? Front. Educ. *4*.

Buchholtz, N.F., Krosanke, N., Orschulik, A.B., and Vorhölter, K. (2018). Combining and integrating formative and summative assessment in mathematics teacher education. ZDM *50*, 715–728.

Dann, R. (2014). Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice. Assess. Educ. Princ. Policy Pract. *21*, 149–166.

Gibbs, G. (1999). Using assessment strategically to change the way students learn. In Assessment Matters in Higher Education: Choosing and Using Diverse Approaches, S.A. Brown, and A. Glasner, eds. (Buckingham [England] ; Philadelphia, PA: Society for Research into Higher Education & Open University Press), p.

Gomez, J., and Rosenberg, L. (2019). In hands of police, facial recognition tech violates civil liberties.

Gross, K., and Bergstrom, C.T. (2019). Contest models highlight inherent inefficiencies of scientific funding competitions. PLOS Biol. *17*, e3000065.

Harb-Wu, K., and Krumer, A. (2017). Choking Under Pressure in Front of a Supportive Audience: Evidence from Professional Biathlon. Univ. St Gallen.

Harwell, D. (2019). Rights group files federal complaint against AI-hiring firm HireVue, citing 'unfair and deceptive' practices.

Hattie, J., and Timperley, H. (2007). The power of feedback. Rev. Educ. Res. *77*, 81–112.

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. J. Pers. Soc. Psychol. *77*, 1121–1134.

Lejk, M., and Wyvill, M. (2001). The Effect of the Inclusion of Selfassessment with Peer Assessment of Contributions to a Group Project: A quantitative study of secret and agreed assessments. Assess. Eval. High. Educ. *26*, 551–561.

Norton, L. (2004). Using assessment criteria as learning criteria: a case study in psychology. Assess. Eval. High. Educ. *29*, 687–702.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science *366*, 447–453.

Penney, J. (2016). Chilling Effects: Online Surveillance and Wikipedia Use (Rochester, NY: Social Science Research Network).

Pope, N. (2001). An Examination of the Use of Peer Rating for Formative Assessment in the Context of the Theory of Consumption Values. Assess. Eval. High. Educ. *26*, 235–246.

Rämö, J., Reinholz, D., Häsä, J., and Lahdenperä, J. (2019). Extreme Apprenticeship: Instructional Change as a Gateway to Systemic Improvement. Innov. High. Educ. *44*, 351–365.

Turunen, J. (2019). The black box problem – what it is and why you should worry about it.

Vihavainen, A., Paksula, M., and Luukkainen, M. (2011). Extreme apprenticeship method in teaching programming for beginners. In Proceedings of the 42nd ACM Technical Symposium on Computer Science Education - SIGCSE '11, (Dallas, TX, USA: ACM Press), p. 93.

Zhang, M. (2013). Contrasting Automated and Human Scoring of Essays. 11.