Pre-course assignment #2

The following thoughts are sort of a rant about how the authors do not stick to good scientific formal practices, but since these came to my mind, here we go:

Rosario Hernández: Does continuous assessment in higher education support student learning?

Its about perceptions of a non-representative sample of university teachers and students of the relationship of assessment and student learning. It focuses especially on the effects of continuous assessment and its formative versus summative role and how university students and teachers experience these effects.

The answer to the question is "almost not at all most of its present implementations". While the necessity of assessment is appreciated by both teachers and students, the potential positive effects of feedback (including continuous assessment) on learning were apparently not materializing enough (feedback not sufficient or not helpful or to late, etc.). The authors offer some advice to close the gap between the possibilities and the reality in supporting student's learning by continuous assessment.

I had mostly negative reactions when reading the paper. Although I agree with the take-home message, the paper is terrible from an academic point of view. Too many non sequiturs.

E.g. I do not see how some of the authors' advice (namely "advocating the engagement of students in managing and monitoring their learning.") can be drawn from their study results as reported in this paper. This advice is most likely a good one, but where is its connection to the study?

The methodology is flawed (OK, they admit it themselves, but that doesn't make it less flawed): If a study with a similar methodology was done in medicine ("Perceptions of a small, non-representative sample of doctors and patients on the effects of drug Y on disease Z"), nobody would take this study serious (at least not here in Finland).

Since I do not have access to the raw data (the actual questions, that were asked), it is very difficult to assess what was really measured. E.g. seemingly there was no differentiation between different types of continuous assessment (peer- versus teacher assessment), which makes the data not very meaningful.

"This study has shown that continuous assessment has the potential to support student learning through feedback and to increase students’ motivation for learning." I agree that continuous assessment has the potential to support student learning (who would disagree?), but this study has not shown what this sentence claims! This study was just analyzing the perceptions of a convenience sample of teachers and students on this topic. Maybe taken together with several other studies, it supports this claim, but this sentence just oversells the actual results that this single study provides.

The author should also take into consideration that this study used self-reporting of a self-selected sample ("lowest grade evidence possible"). Hence the true numbers are likely to be even more dire! 16.5% of students reported to have acted upon the teachers feedback? What does "evidence" mean is this respect? How did the authors get evidence for the students' claims? The real number might be 5%, which paints a very bleak picture of the real situation of how much feedback is helping students! I light of this situation, I find it remarkable and interesting that students seem to be more uncritical than the teachers in their fundamentally positive attitude towards the necessity of assessment for student learning.

Another thing that surprised me when looking at all the articles that were for choice for this assignment is that these journals seem not to have limits on the article length. These articles are all massively longer than most of what I have published. Maybe here again an example that assessment criteria influence the performance: Many universities did count page numbers (and some still do) in the evaluation of research output. Godhard's law was again at work: When measure becomes a target, it ceases to be a good measure. But if this is true, student assessment is like to squaring the circle, because students always will aim at good grades...

Lucy Johnston and Lynden Miles: Assessing contributions to group assignments

The take-home message is that self- and peer evaluations can be used in the assessment of individual contributions to group work. The authors also tried to distinguish between different models to include the self- and peer evaluations into the final grade. Pros and cons for e.g. ex- and including self-assessment or whether to apply the results of the self- and peer assessment as a modifier of the written assignment or an independent factor for the final grade were discussed, but no definitive answer was reached although the authors lean towards using peer assessment in a modifier role for grading, mainly because because students rated their own contribution differently then their peers did, who supposedly evaluated accurately the contribution of others to group work. They also discuss in length the possible gaming strategies that students might have used to tip the grading into their favor, but conclude in the end that none of this was happening and if, it was not significantly influencing the grading. "We interpret these data as indicative of the students taking the contributions rating task seriously." A possible alternative explanation: the students out-witted the teachers by not letting them get insight into their motives and strategies for evaluation.

Even though the topic is interesting (that's why I chose this paper), I could hardly get myself reading to the end because this paper is terrible! It is written in a very unclear style and could be made much easier to understand if the authors would have just tried harder. It's a failure of peer review (which ironically is the topic of the paper). However, I am not sure whether the authors themselves fully understand their own calculations. The statistical methods of the paper are clearly flawed. When the authors allow for different group sizes, they start to compare apples and oranges. They find differences when they compare the different outcomes with respect to the group size, but their immediate reaction should have been to stick to one group size!!! They could have easily eliminated 4 of the 5 groups which were not n=4 by re-assigning students! It's a basic rule of science not to modify too many factors because this obfuscates the results and makes them very difficult to interpret.

They have done multiple comparisons and their p-values have not been adjusted correspondingly (or the authors do not mention it, which is almost equally atrocious). On the other hand their numbers in some comparisons are so low that they would not allow to detect real differences, even if such did exist! Also the fact that the groups were self-assembled is a limiting factor, all groups should have been randomly assigned in 4 students/group.

"There was no correlation between self- and peer-assessments, with self-assessments being higher, on average, than peer-assessments." This sentence contradicts itself! If self-assignments are on average higher, then there is a correlation (even if it's a weak one). If the difference is not statistically significant, why do they talk about it?

The math doesn't make sense to me also at several other places: There seem to be simple arithmetic mistakes in Table 2, e.g. the in the median column 13.65 - 13.41 ≠ 0.138. There are several other places in the table, where I cannot follow where the numbers are coming from. Or did the authors totally mislabel the Table 2: Is it the "marks" or the "change of the marks" what they are showing here? Column 1 indicates that it's the marks, but the other columns do not support this notion.

If such math already is faulty (or at least not documented), how am I supposed to trust the rest? If the students consistently evaluated themselves higher than the group did, then there should have been a correlation; but the authors claim that there was none. There would be no correlation if students evaluated themselves randomly. However, I have no time to replicate their math and since the authors do not make their raw data available, it might not be possible to figure out what and how they actually got to the numbers that they present.

I think the authors also calculate and discuss irrelevant numbers. They talk e.g. about the "mean change of marks". What is this supposed to reflect? If exactly 50% of students have a better grade and 50% a lower grade post assessment the "man change of marks is 0, but this doesn't mean that there is no influence on the grading. What's the purpose of such measure? The authors state in the next sentence exactly the same: "There was, however, a wide range of changes to marks, from a decrease of 22.12% to an increase of 20.18%". Why do they at all bring up the useless measure of "mean change of marks", which just confuses the reader and inflates unnecessarily the text? the important thing is that the grades change and that the change can be substantial.

The Dunning-Kruger effect was NOT seen! But since I do not trust their calculation abilities, I do not trust that they looked for it correctly! They themselves remark, that this results contradicts previous research. maybe they should have taken this as a reason to double-check their numbers! "However, where self- and peer-assessments have differed it appears as if students actually tended to assess themselves lower than did their peers (Krause & Popovich, 1996), especially the more able students (Lejk & Wyvill, 2001)."

"It has been suggested that self-assessment varies as a function of ability level, with higher ability individuals under-rating themselves and lower ability individuals over-rating themselves (Lejk & Wyvill, 2001)."

About the language they use:

Are these authors really that insensitive??? How does this get through peer review? "Sex" has many meanings and it is inexcusable not to replace it by a more accurate and appropriate description! Words do matter! "mixed sex work groups", => mixed gender work groups, "the sex of the group": I know what they want to say, but they formulate inaccurately. A group does not have a sex! I might be biased, but sloppy writing might reflect sloppy thinking!

The language check did not quite catch all mistakes (which is remarkable for English native speakers): "Discrepancy scores for those in the middle quartiles feel [sic!] in between the highest and lowest quartiles."

Video #1: Great news for my son since he wants to study math at HY! really: This video sounds almost like utopia: too good to be true. I wonder what would happen if I proposed to abolish exams for some of our study program's more important courses. Maybe I should just do that and see what happens. In a faculty where the workload is already close to maximal for everybody, how can one implement such radical changes? For me personally, it's clear that this is the right direction to go, but where do we get the additional time? Johanna mentioned the the few problematic cases were those, where there was not enough contact between her and the students. Such teaching takes MORE time and if there is anything that is important nowadays at the university (to keep the ministry's money flowing with the new allocation model), than that is student throughput... Please answer: How many of the courses (in total and %) at the math department in Kumpula are organized this way?

Video #2: The video seems incomplete, an excerpt of something bigger. Simply three examples which show that course grade does not necessarily reflect whether or not learning and teaching goals were reached. However, if it was possible to get a grading of 5 without any deeper understanding of the topic (example 2) simply by learning by heart, then something should be changed in the way the evaluation is done. Example 3, on the other hand, might need some assistance related to his time management. Unfortunately, the video tells nothing about what level the students were (1st, 2nd, 3rd year students?) and what level of time management skills can or cannot be expected from them and what support they received (or did not receive) from university to learn those skills. Compared to video 1, please note the difference in student numbers for the course! Hence, therefore perhaps the traditional lecture format with its drawbacks.

How should group and reflection assignments be weighted in the final grade?

The answer depends of course on which task you want the students to spend more time and effort on. For me individually (being equally challenged by both tasks), 50-50 will perhaps work best to avoid justifying to neglect one over the other task. Demoting the group task for the simple reason that one has not as much control over it seems only superficially OK. It's about group work and group work is reality for most of us (including teaching and grading): we control only part of the final outcome of our work as we always are part of a bigger entity and always have to rely on others to do our work.

Individual weightings in the group assignment: same grade for everyone or is the individual's participation in the group work taken into account e.g. through peer evaluation? (For this kind of peer evaluation, see Team Q model in a separate folder.)

In the beginning I thought: Of course according to participation! Otherwise, I expect there to be minimal participation from some members (perhaps including me). However, now, about a week later (and after reading some of the course material like "Lucy Johnston and Lynden Miles: Assessing contributions to group assignments" and the Team Q description), I start to have doubts. In the end, the answer depends on how participation is exactly measured. Peer group assessment (with or without self-assessment) appears an option, but the exact procedure would matter. One can imagine a scheme, where every group member can have a rating between minimal and maximal amount of points (like the Team Q example), but one can also imagine a scheme where a certain number of points are distributed among the group members. These two options might lead to very different dynamics inside the group. Since the Team Q M$-Word document does display terribly when opened with LibreOffice, I would give all group members the same grade. The rationale being that my time is limited and it would take lots of contemplating to give just and justified ratings to the other group members. BTW: Who uses M$-Word these days to make such evaluation sheets? It's simply not the right tool for the purpose, but since in the Team Q article, they use "some statistical software, that is not compatible with the numeric zero", this is probably the smaller problem. Why do they not disclose the software they use? This should be science and not a black box and the software that is used to do the analysis should be always disclosed under "Materials & Methods". Just recently, "some software" was discovered to have bugs, which apparently resulted in many wrong results which have been published (the "Willoughby-Hoye" scripts to be specific).