Hive Helsinki technical assignment, Bocal internship   Esa Halmkrona, ehalmkro

Assessing the quality of peer-to-peer evaluations, an MVP

For my assignment, I had to make a script to find students who have done 'bad evaluations' during a certain timeframe. For the technical part of the assignment I decided to follow the minimum viable product process by starting out by finding the most obvious bad evaluations. The internal *Art of Peer-Evaluation* document by the 42 pedagogical team lists some features of a 'good' evaluation, including:

- taking your time
- following the Norm
- exchanging ideas
- being constructive

Additionally, the document emphasizes the role of feedbacks from both the evaluator and the evaluated person for showing the quality of the defense session.

For my naive first definition of a bad evaluation I decided to just use the opposite of a good evaluation. A bad evaluation is rushed, not following the evaluation form/Norm/subject**,** has very little debate or conversation and is not constructive leading to bad review by the evaluated. Many of the characteristics of a bad evaluation are difficult to catch technically. My main focus with the technical MVP was catching the ones that are the most obvious. I use a simple scale to measure the so called quality (or 'goodness') of an eval with the following parameters:

- length of the eval session
- length of the text feedback given by the evaluator
- feedback of the evaluation session, given by the evaluated person

For the MVP, these three factors add up to a single number, evaluation quality, that I measure in simple percentages. An evaluation's quality can thus vary from 0% to 100%.

Testing my script and seeing its results on actual data, I quickly found out the data is very interesting, not in singular cases but when looking at a bigger student population. For example, when comparing Hive Helsinki evaluations to all the 42 schools, there are some differences even with the very naive, non-detailed and low-variance way of measuring evaluation quality.

Ideas for improvement

For the tool to work in real-life setting, it should be pivoted towards being a sort of an everyday watchdog, alerting when a clearly bad evaluation was submitted. It also could give out some historical data on how for example piscines affect the evaluations. These data could be used to direct attention where it might be needed. 42 API has a lot of data on the evaluations, projects and the people involved. This could be used to e.g.

- filter keywords in feedback
- check that the feedback was written in English
- compare the evaluations done remotely and at school
- run ‘shadow moulinette’ to see if the evaluators’ grades match up with the actual test results
- make ‘evaluation profiles’ of evaluators to guide the evaluation matching

Technical solutions for pedagogical problems?

An evaluation’s quality is something that can not be accurately measured using API data. Thus a technical solution for the problem is at best very partial. A bad evaluation in a peer-to-peer learning context is a situation where the amount of learning is affected by the actions (or lack of actions) of either the evaluator or the evaluated. While using technical solutions to help find these situations and try to measure the overall evaluation experience at the school is definitely possible, the actual solution needs to be pedagogical.

As Hive Helsinki is a peer-to-peer learning environment, learning within Hive and within the 42 curriculum happens in a strongly cultural context. Hive is a learning community. The possible situation of having bad evaluations happen *en masse* would be a cultural issue. Aside from the single bad evaluation here and there, addressing the issue would be changing the culture regarding evaluations.

I believe we are at a great position now at Hive Helsinki, the school is new and the culture is still forming. It seems we are already doing well regarding the quality of evaluations and to further support the cultural development would not require too much action in terms of pedagogical interventions. We don’t need a cultural shift, just pedagogical supervision. My script, when further developed, could work as one technical tool to aid with this supervision. The possible pedagogical intervention when finding a pattern of lessening evaluation quality could be a simple talk with the students involved.

More broadly, forming and directing the learning culture requires working together with the student body (not to forget the role of the student board as a mediator) in order to scaffold and direct the students’ learning. One important factor of quality peer-assessment is having clear criteria on what is a good project. The criteria needs to be clear but there also needs to be room for discussion and debate regarding the different approaches. Most of the time a bad evaluation is not the fault of the evaluator or the evaluated person. Instead, a bad evaluation happens because of unclear instructions, time constraints or too loose of an evaluation culture. While we can’t affect the outcome of single individual evaluation sessions, I see great possibilities in building and maintaining a solid, rigorous but polite evaluation culture while the school still is at its infancy. While it is not possible to get involved students' learning process directly in a completely peer-to-peer school, I think the purpose of the pedagogical team is to set up an environment and a culture where learning happens most efficiently. Technical tools would lessen the everyday load of pedagogical supervision freeing resources for the more meaningful pedagogical action.